CS4740: Introduction to NLP
Project 4: Deception Detection

Part 1 is due: 1:25 PM, Thursday, April 18, 2013
Final report is due 1:25 PM, Tuesday, April 30, 2013

1 Introduction

In this project, you are to implement algorithms to find deceptive reviews. This is in follow up of the paper [Ott et al., 2011] you recently read in class[1]. As you must have noticed in the paper that simple features like unigrams and bigrams do very well in discriminating true vs fake reviews. More surprisingly, humans are very poor at this task. The motivation behind this project is to find answers to the following questions:

1. How can humans, intelligently, write deceptive reviews which are hard to classify using machine learning techniques.
2. How can humans boost their own performance in discriminating the two kinds of reviews.
3. How will the machine learning algorithms perform on such reviews.

As you must have sensed by now, above questions are potential research directions to which no one have definite answers. Our aim here is to explore these questions further through various experiments. It is important to note here, that answers to these questions are largely domain dependent. In this project our domain will be hotel review dataset which is generously made available to us by Myle Ott. This project will proceed in a number of phases (described below) during which you will try answering each of the above question.

2 Different phases of project

This project will proceed in the following four phases:

1. **Phase 1:** In order to help you write reviews which are difficult for machines to classify we will seed you with those deceptive reviews which were wrongly labelled as truthful by human judges in [Ott et al., 2011]. We will also provide you with some truthful reviews, so that you could compare the two kinds of reviews. You should go through these reviews and identify the

[1] This project will give you an insight into coming up with research ideas while writing critique

features which differentiate deceptive and truthful reviews. Note that, the features you identify might not be possible to code, these features are only meant for humans to distinguish the two kinds of reviews. You should write down as many features (observations) as possible and upload them on CMS and **submit hard copy in class**.

2. **Phase 2:** After you have completed phase 1, we will upload fresh unlabelled data. Your job will now be to manually classify these reviews into two classes. Since you will be working in groups it would be interesting to observe how different members of the groups classify a given review. As a group you are supposed to submit just one label per review. Your group can do majority voting among its members to choose a label. We will launch a kaggle competition for this phase.

3. Phase 1 and 2 will constitute the **part 1** of the project. Since part 1 doesn't have any coding exercise we hope all the groups will do well in kaggle competition. After the kaggle competition is over, we will release the gold standard labels. You are expected to identify the reviews which were misclassified by you and comment on some of them in the final report. Please note, we are more interested in a general theory as opposed to case by case theory.

4. **Phase 3:** After phase 1 & 2 you should (hopefully) be doing well is distinguishing deceptive and truthful hotel reviews. By now you should also have some insight into why features like unigram and bigram perform so well in classifying reviews. Phase 3 is one of the most important part of this study and "Just" for this phase you will not work in a group i.e. you will work alone. For this phase you will write **two deceptive reviews** (one positive and one negative) and upload them on CMS (only). Instructions on writing these reviews are given in section 4. Make sure you read the instruction very well and also follow them. We will go through the reviews submitted by you to make sure they follow the guidelines.

5. **Phase 4:** This is the final and most open ended phase of this project. The data collected from phase 3 will be released via CMS along with the data we already have with us. Now your goal in this phase is to design features and use any machine learning algorithm to classify the reviews as deceptive or truthful. We will launch a kaggle competition for this phase as well. In the following sections we will describe this phase and our expectations from you in more detail.

3 Data sets description

3.1 Phase 1 data:

Phase 1 data consists of 40 reviews which were labelled by three human judges. The format of the data is as follows:
IsTruthful;review;author1;author2;author3

- **IsTruthful** - corresponds to the true label of the review. Deceptive reviews are labelled as 0 and truthful reviews as 1.
- **review** - corresponds to the review of the hotel.
- **authori** - corresponds to the label by i^{th} human judge.

In order to avoid any kind of bias we have removed hotel names from this data.
You should very carefully do the following study on this data:

- Try to understand the reasons behind poor human performance.
- Extract features which a human can use to differentiate the two kinds of reviews.
- How can humans write good deceptive reviews which are difficult to recognize using machine learning algorithms.

What to submit from phase 1?

- Submit a preliminary version (maximum 1 page in length) of above study online on CMS and printout in class on April 18, 1:25PM.
- In your final report you will discuss this study in greater detail. By then you would have seen much more data so you could revise your hypothesis as well.

3.2 Phase 2 data:

After you have submitted the phase 1 report we will release the phase 2 data and launch a kaggle competition. Phase 2 data consists of 120 reviews and has the following format:
IsTruthful;review

- **IsTruthful** - this filed has ? and you should fill it with either 0 or 1.
- **review** - corresponds to the reviews.

In this phase your job will be to manually label the reviews as deceptive (label 0) or truthful (label 1) based on the experience you gained during phase 1. It would be

interesting to observe how different members of a group perceive a review. Every group will submit one label per review. After the kaggle competition is over (on April 21, 1:25PM) we will release gold standard labels for this data.
What to submit from phase 2? Using the gold standard labels, you should carefully observe those examples which were misclassified by you. In your final report you should discuss and analyse some of these examples and citing possible reasons for misclassification. You should also discuss in the final report how your group reached consensus in situations of differing opinions. It would be advantageous to observe the examples misclassified by you before proceeding to phase 3, this would help you to perform better in the next phase.

3.3 Phase 3 data:

This phase will start after phase 2 kaggle competition is over. In this phase you will work individually and will write two deceptive reviews (one positive and one negative) and submit them on CMS by April 24, 1:25PM. You should write reviews with the aim of making them as close as possible to true reviews. Read section 4 very carefully for detailed instructions on submitting the reviews.

3.4 Phase 4 data:

Data for this phase will be released after phase 3 data is processed and graded. We will mix the reviews collected from phase 3 with the data we already have. Below is the data set statistics, unknowns will be updated before releasing the data.

Training data: Consists of XXX reviews out of which PPP are positive reviews about the hotel and NNN are negative. Positive and negative reviews are further divided into deceptive and truthful class.

- Positive reviews have DDD deceptive and TTT truthful reviews.
- Negative reviews have DDD deceptive and TTT truthful reviews.

Validation data: Consists of XXX reviews out of which PPP are positive reviews about the hotel and NNN are negative. Positive and negative reviews are further divided into deceptive and truthful class.

- Positive reviews have DDD deceptive and TTT truthful reviews.
- Negative reviews have DDD deceptive and TTT truthful reviews.

Test data: Consists of XXX reviews.
In section 5 we list a set of studies which every group must do on this data. We also provide a list of possible extensions. Please note, we will launch a kaggle competition for this phase as well.

4 Instructions on submitting the data for phase 3

Every student will be assigned a hotel-url in the comments filed in CMS. You should submit *exactly* two deceptive reviews electronically on CMS by April 24, 1:25PM. **We will create separate submission page for each review**. Following are the instructions for writing the deceptive reviews:

1. You have to submit two reviews (one positive and one negative) for the same hotel and make sure you submit them on the correct CMS submission page, else you will loose marks.
2. Make sure you write review for the correct hotel.
3. Reviews should be submitted in plain text and should have the following format: $hotelname - review$
4. Students should individually write their reviews i.e. no collaboration.
5. **The reviews should focus on spatial configuration of hotel** i.e. its room, bathroom, food, location etc. This is very important because this would make the reviews closer to true reviews. You should also incorporate some of the features you discovered during phase 1 and 2. Please try to be innovative.
6. Each review should atleast be 150 characters long excluding white spaces and should **not contain misspelled words**.
7. Obviously, reviews should not be plagiarized.

We will go through the reviews to make sure they meet the instructions.

5 Phase 4 Experiments

In this phase you are to design various features and implement algorithms to classify reviews. Every group must experiment with atleast three different features and two different learning algorithm. Below we give a list of possible features and learning algorithms.

5.1 Features

[Ott et al., 2011] describes a bunch of features which do well in deception detection. Below we list some features which you could try:

- Word n-gram.
- Character n-gram.
- Part of speech tags[2].

[2]You may use any POS library

- Term frequency-inverse document frequency. (Refer:Wikipage)

Instructions:

- You must implement character n-gram features and two other features of your choice.
- You must implement the code for extracting features (exception POS based features). You may use any tokenizer.
- If you plan to implement features outside those stated above and in [Ott et al., 2011] then have them checked with teaching assistants.
- For n-gram based features you should experiment with different values of n.

5.2 Learning algorithms

You should experiment with atleast two learning algorithms. Below we suggest few possible algorithms, but you are free to try algorithms not listed below:

- Support Vector Machines (SVM): Two popular implementation of SVM are SVM-*light*[3] and LIBSVM[4]. Whichever library you use, you should cross-validated to choose the best hyper parameter (C). Generally the following is a good range for the values of C $\{10^{-3}, 10^{-1}, 10, 10^2, 10^3\}$.
- Language models: You can develop two language models, one for deceptive and one for truthful reviews, and use them to classify a new review based on its perplexity value.
- k-nearest neighbour (kNN): The input to this algorithm is a user defined number, k. It proceeds as follows: Given a test data point, its distance in feature space is calculated from every training data point. Top-k closest training data points are then retrieved based on the distances calculated above. The test data point is then assigned the label of the class which forms a majority among those top-k points. For more details on this algorithm you can refer to its wiki page. The distance between two data points can be defined in many ways, for the purpose of this project you can use the euclidean distance as a metric. You should cross-validate over different values of parameter k.

5.3 Extension:

Below we list some possible extensions. Every group must do atleast one extension.

[3] http://svmlight.joachims.org/

[4] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

- Implementing an additional feature or learning algorithm. While many groups might choose to do this, we expect extensions not to be very straightforward.
- Extending kNN to use a distance metric more suitable for text categorization e.g. hamming distance.
- Using polarity of the review as a feature in deception detection. For this you can cascade two classifiers. First classifier will assign a polarity to a review i.e. positive or negative. While the second classifier will use polarity as one of the features and classify them as deceptive or truthful.

5.4 What to report from phase 4?

You should report and discuss the performance on validation and test set for the algorithms, features and extensions you implemented . You should also plot the learning curves i.e. you should vary the number of training points and train a classifier everytime and evaluate it on validation set. You should then plot on x-axis the number of training points and on y-axis the performance on validation set. You should plot learning curves for different algorithms and features. Make sure yours plots are legible (not too cluttered), have legends and required labelling. You should not put too many separate plots (i.e. try overlaying relevant plots), otherwise it becomes hard to derive any meaningful information from them. Discuss in the final report any trends you observe in these plots.

6 Complete timeline and points distribution

- Phase 1 (10 points):
    1. Data will be released on CMS on April 13.
    2. One page summary of your study will be due online on CMS **as well as hardcopy in class** on April 18, 1:25PM.
- Phase 2 (20 points):
    1. Data for this phase will be released following the submission of phase 1 and kaggle competition will also be launched.
    2. Kaggle competition ends on April 21, 1:25PM.
- Phase 3 (10 points)[5]
    1. Every student will be assigned a hotel-url via CMS following the end of phase 2.

[5]Due to this phase, students within a group can get different final scores.

2. Review submission will be due online on CMS (only) on April 24, 1:25PM.

- Phase 4 (80 points):

1. Data for this phase will be released following the end of phase 3 and kaggle competition will also be launched .
2. Final report will be due online on CMS **as well as hardcopy in class** on April 30, 1:25PM.
3. Kaggle competition will carry 20 points.
4. Extension will carry 20 points (implementation + evaluation).
5. Implementation of algorithms, features and their evaluation and discussion will carry 30 points.
6. Other discussion e.g. comparison with gold standard etc. will carry 10 points.

Final report will have contribution from all the four phases. Follows the instructions in subsections/paragraphs titled "What to submit/report". Report should not be more than 5-6 pages in length.

Advice: Form a group as early as possible (**advisable size 4-6 members**) and start with implementing phase 4. You should try to finish the implementation before we release the data for final phase. This would allow you to get started with experiments as soon as the data is released.

References

M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1032`.