|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | Native language binary classification | |  | |  | |
|  |
|  |

Abstract

Most of the world’s population is bi-language. Moreover, even though the dominant language in the internet (particularly in social media) is English, there are evidences that most of the dynamic content is created by non-native English speakers. That is why the problem of distinguishing between native and non-native speakers is drawing attention. Potential applications of this task are: Teaching English more efficiently, identifying target audience based on native language etc.  
In this work we took an approach that is content independent - we are modelling text by the function words which occur in it. We then use machine learning techniques to distinguish between native and non-native English speakers, yielding solid results.

Introduction

Native language classification is a well studied problem. One popular approach is Bag of Words (BOW). In BOW the text is represented by a feature vector of dimension N (N is the size of the vocabulary or a subset of the vocabulary). Each dimension is the count of the corresponding word form the vocabulary occurring in the text. Usually stop words are excluded from the feature vector. This approach (as will be shown in the results section of this document). However, this approach suffers from a main disadvantage which is content dependency. For instance, if a classification model is trained on specific domain corpus (e.g. Sports, Polticis, Traveling etc) this model will consider words from the specific domain with much higher importance compared to words out of the specific domain. This presents an issue in generalization across domain. Also there is a time relevance issue – domain dominant words can change significantly over time (this is particularly noticeable domains such as politics and sports – where things changes rapidly). This could lead to poor classification results that might force retraining of the model for the new domain or the new period.

In order to overcome the content dependency issue we took a content independent approach by observing the use of function words (which do not carry content) for native and non-native English speakers. Because function words are not domain nor period dependent - this approach is robust to the issues described above and yields firm results for binary classification.

We use 2 different data sets for native and non-native English speakers respectively. Native language data set was extracted from Reddit – a popular American discussion website. We had 6 different countries of origin for native English speakers (USA, UK, Ireland, New-Zealand, Canada and Australia) for each we used an equal portion in this work. Non-native English speakers dataset was taken from TOFEL (Test Of English As A Foreign Language) – a collection of assays written by non-native English speakers as a university entrance test. The country of origin of the writer was not speficied. We used function words dictionary from the academic resources of Sequence Publishing as a base, and extended it manually during our training process.  
We used several well known classifiers (SVM, Decision Tree and Naive Base) with 3-fold cross validation for this task and compared the results of the function words approach vs. BOW approach.

Related Work

The idea of collecting annotations from volunteer contributors has been used for a variety of tasks. Luis von Ahn pioneered the collection of data via online annotation tasks in the form of games, including the ESPGame for labeling images (von Ahn and Dabbish, 2004) and Verbosity for annotating word relations (von Ahn et al., 2006). The Open Mind Initiative (Stork, 1999) has taken a similar approach, attempting to make such tasks as annotating word sense (Chklovski and Mihalcea, 2002) and commonsense word relations (Singh, 2002) sufﬁciently “easy and fun” to entice users into freely labeling data.

There have been an increasing number of experiments using Mechanical Turk for annotation. In workers provided annotations for the tasks of hotel name entity resolution and attribute extraction of age, product brand, and product mode. The annotations found to have high accuracy compared to gold-standard labels. Kittur et al. (2008) compared MTurk evaluations of Wikipedia article quality against experts, ﬁnding validation tests were important to ensure good results.

In general, volunteer-supplied or MTurk-supplied data is more plentiful but noisier than expert data. It is powerful because independent annotations can be aggregated to achieve high reliability. Sheng et al. (2008) explore several methods for using many noisy labels to create labeled data, how to choose which examples should get more labels, and how to include labels’ uncertainty information when training classiﬁers. Since we focus on empirically validating MTurk as a data source, we tend to stick to simple aggregation methods.

MTurk - Overview

MTurk functions as a one-stop shop for getting work done, bringing together the people and tools that enable task creation, labor recruitment, compensation, and data collection. The site boasts a large, diverse workforce consisting of over 500,000 users from over 190 countries who complete tens of thousands of tasks. Individuals register as ‘‘requesters’’ (task creators) or ‘‘workers’’ (paid task completers). Requesters can create and post virtually any task that can be done at a computer (i.e., surveys, experiments, writing, etc.) using simple templates or technical scripts or linking workers to external online survey tools (e.g., SurveyMonkey). Workers can browse available tasks and are paid upon successful completion of each task. Requesters can refuse payment for subpar work. Being refused payment has negative consequences for workers because requesters can limit their tasks to workers with low refusal rates.

Project Build Process

We used tweets from Twitter to obtain data for the HITs (Human Intelligence Tasks). We examined the information collected from the questionnaires that were answered by individuals and not by organizations. The information we extracted from Twitter about the diseases also removed tweets that contained words that do not indicate diseases (inhaler, band-aid, etc.). These tweets were deleted and not entered to the questionnaires (HITs).

Before we implemented the project, we ran several small experimental projects. There were several purposes for running the mock projects:

1. Learn how the MTurk system works.
2. Examining the response time of people who answered the questionnaires.
3. Examine the response and professionalism of people when meeting texts from Twitter.

After analyzing the results of said projects, we ran the main project - distributing a large group of questions with the tweets.

Task Design

In general, we follow a few simple design principles: we attempt to keep our task descriptions as succinct as possible, and we attempt to give demonstrative examples for each class wherever possible. We have restricted our study to tasks where we require only a multiple-choice response. For every task, we collect ten independent annotations for each unique item.

Project Build Process

To set up a project in Amazon's system, a template needs to be selected out of several possibilities. We created a general questionnaire that was intended for people who met our conditions. The format of the questionnaire also required the following:

* A definition of the payment for each worker.
* The response time for completing each questionnaire.
* The total number of questions.

For creating the questionnaires, two other main things are defined:

1. A questionnaire format that contains a fixed text with variables that are received in each variable information questionnaire (**Appendix A**).
2. A CSV file in which each row contains information for a HIT - input as variables in the questionnaire, ordered by column name in the CSV file (**Appendix B**).

Dataset

We collected 980 tweets that mentioned the following diseases: HIV, Asthma and Fibromyalgia. Each disease was divided into 3 categories: Self or family, someone else or general. Each tweet was labeled by 3 experts who determined the categorization of the tweet – gold label. The number of tweets collected is:

* HIV:
* Self or family - 94 tweets.
* Someone else - 168 tweets.
* General - 200 tweets.
* Asthma:
* Self or family - 199 tweets.
  + Someone else - 2 tweets.
  + General - 95 tweets.
* Fibromyalgia:
  + Self or family - 200 tweets.
  + Someone else - 22 tweets.
  + General - 0 tweets.

Source of data

The data that we used in the learning and testing process was based on tweets from Twitter.

The reasons for using this data were:

1. We were interested in the study of stigmas in social networks.

2. Twitter has an interface that allows to collect tweets.

3. The tweets have limited number of characters.

4. Any user or group is trackable for future research.

Fleiss' Kappa

Fleiss' kappa is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. This contrasts with other kappa's such as Cohen's kappa, which only work when assessing the agreement between not more than two raters or the interrater reliability for one appraiser versus them self. The measure calculates the degree of agreement in classification over that which would be expected by chance. As this large and positive number indicates a better agreement. On the other hand, a negative number indicates poor agreement.

Tweet Annotation Label

In this experiment, 980 tweets were selected from the gold label dataset, and 3 affected annotations were collected for each of the three label types. The results contained a total of 2940 affected labels. Fleiss' kappa was calculated for each labeling of a worker. In addition, we presented each sub-category of the disease its Fleiss kappa score.

|  |  |  |  |
| --- | --- | --- | --- |
| Fibromyalgia | Asthma | HIV | Annotations |
| 0.305 | 0.336 | 0.133 | Self or  family |
| -0.028 | 0.075 | 0.245 | Someone else |
| -0.03 | -0.052 | -0.133 | General |

Table 1: Calculate Fleiss' kappa by categories

The results in Table 1 shows that the agreement score between the workers was poor. This could have happened for several reasons:

1. The given tweets were incomprehensible.
2. The choices were confusing, so there could have been a situation where there were very small differences between the 2 answers.
3. Selected workers were not qualified or were not experts on Twitter.
4. The number of questions in that category was too low.

As a result, we were interested whether the general rating agreement for each disease will show improvement, or not.

During the data collection, we focused on three diseases that represent different levels of stigmatization.

|  |  |  |  |
| --- | --- | --- | --- |
| Fibromyalgia | Asthma | HIV | Annotations |
| 0.376 | 0.418 | 0.25 | Kappa score |
| 296 | 222 | 462 | Tweets count |

Table 2: Calculate Fleiss' kappa score for each disease

The results in Table 2 show that the agreement score has improved and in some cases, became positive. This testified, as we hypothesized, that there is a connection between the number of tweets for the same disease to the Kappa score. The agreement score amongst workers in the asthma and fibromyalgia categories is higher than in their sub-categories. Therefore, to estimate the differences, it was necessary to add additional tweets to each sub-category, and thus, the agreement in the sub-category will increase just as in the category itself.

Labeling

In this experiment, we checked if we could identify a label as a gold label even though the Fleiss' kappa scores were low. We wanted to check the quality of the labeling even though the Fleiss' kappa scores were low or negative

|  |  |  |  |
| --- | --- | --- | --- |
| Fibromyalgia | Asthma | HIV | Annotations |
| 170 | 137 | 193 | Full consent |
| 113 | 79 | 256 | Consent |
| 13 | 6 | 13 | Disagree |
| 296 | 222 | 484 | Total |

Table 3: The agreement on labeling between workers

The results of Table 3 show that, despite the Fleiss' kappa score being low, for most of the labels there was a very large consensus amongst the workers. This indicates that Fleiss' kappa is effected by small disturbances, meaning that the smallest discrepancy between workers reduces the Fleiss' kappa score. The next thing that needed to be checked was the quality of a worker’s product compared to labels made by experts.

The results of Table 4 show that in most cases the workers agreed with the experts. For each disease, the accuracy percentages were:

* HIV:
  + Self or family:
  + Someone else:
  + General:
* Asthma:
  + Self or family:
  + Someone else:
  + General:
* Fibromyalgia:
  + Self or family:
  + Someone else:
  + General:

The percentages shown above are the number of tweets that were correctly classified by the workers for each category versus the number of tweets that were classified by the experts for each category.

An analysis of the data shown in Table 5 shows that, for most of the categories, there was a very good percentage of accuracy. On the other hand, if we look at the part where the workers and experts disagreed, we can see high values. This might be explained by the inaccuracy of the category definitions, which could explain the high percentage of accuracy of true positives in each category and the high percentage of false positives.

|  |  |  |  |
| --- | --- | --- | --- |
| Disagree | Consent | Full Consent |  |
| 5 | 12 | 12 | HIV - Self or family |
| 99 | 19 | 138 | HIV - Someone else |
| 30 | 124 | 10 | HIV - General |
| 7 | 28 | 118 | Asthma- Self or family |
| 7 | 5 | 2 | Asthma - Someone else |
| 49 | 0 | 0 | Asthma - General |
| 18 | 30 | 155 | Fibromyalgia- Self or family |
| 4 | 0 | 2 | Fibromyalgia - Someone else |
| 7 | 61 | 6 | Fibromyalgia - General |

Table 4: Comparison of the gold labels made by experts versus workers. The comparison between the gold label and the label determined by the workers (after a majority decision).

Work Quality of the Worker

In this section, we wanted to examine the quality of the worker's product according to the following measures:

* Percentage of accuracy.
* Amount of HITs.
* Average time to solve all the HITs.

Accuracy vs HITs

In this experiment, we examined the quality of the workers' product according to the amount of HITs they solved. We wanted to check whether there was a correlation between the amount of HITs that a worker solved and his accuracy percentage. According to the results, presented in Figure 1, we did not find a connection between the HITs count and the accuracy. We saw that there were a lot of workers who did little work and had a low percentage of accuracy and on the other hand, there were workers who did more work with a very high percentage of accuracy.

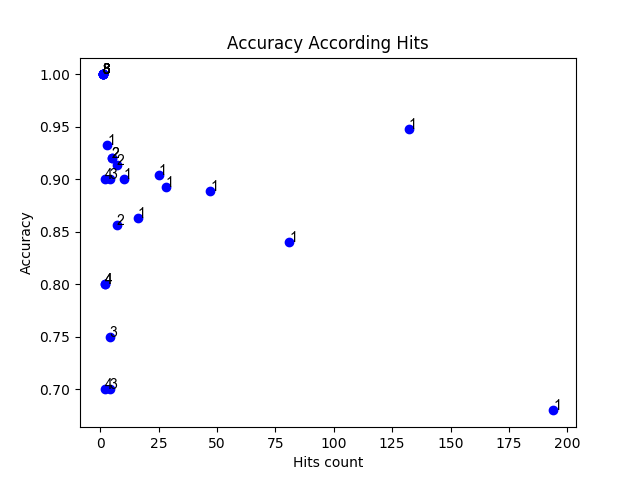


Figure 1: Labeling disease by categories. The numbers above the points indicates the number of workers who have the same percentage of accuracy.

Worker Average Time

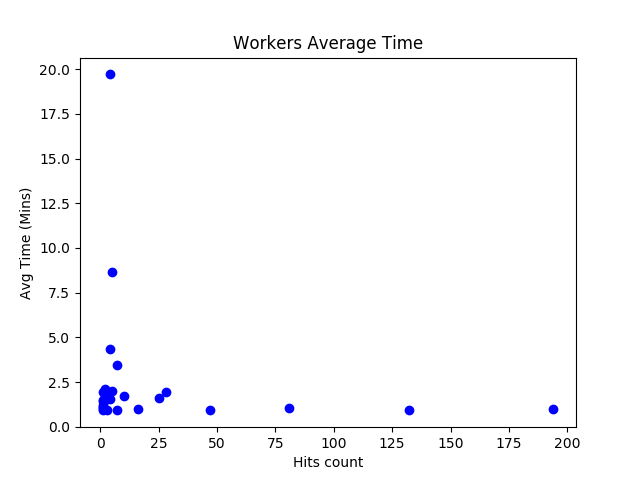
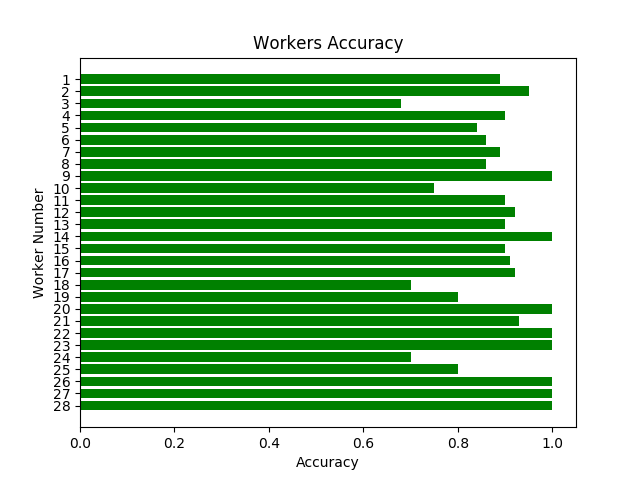


Figure 2: Average time worked by the amount of HITs that he did.

In this experiment, we wanted to examine how many minutes the worker spent on solving the HIT. With the results shown in Figure 1, we checked if there was a direct correlation between the time a worker spent on his work and its quality. In addition, we wanted to see the average time it took to solve a HIT.

According to the results shown in Figure 2, it took an average of 72 seconds for each worker to solve a single HIT. In addition, the results for each HIT took a few minutes to solve, indicating that the task was not difficult to understand. Finally, following the results presented in Table 4 and Figure 5, we concluded that the differences between each category were too minuscule. As a result, we received a very high number of false positives.

Worker Average Time

 Figure 3: The quality of work of the workers among themselves.

In this experiment, we wanted to examine the percentage of accuracy amongst the workers themselves. We wanted to know which workers achieved a high rate of accuracy so that in the future we would be able to work with them again and give them a larger compensation to insure their future participation, if necessary. As for the workers who had achieved a low accuracy percentage, we would be able to block them in the system and prevent future participation in solving our next HITs.

Extra

In this experiment, we will show the results for 2 questions we asked in each HIT for which we did not have gold labels. These questions were asked to help us examine the results obtained and check the quality of the workers' work. For any disease label (Self or family, someone else or general), We asked the following questions:

* Does the tweet talk about disease?
* Is the tweet sarcastic?

The first question examined the quality of the worker's work. If he answered that the tweet was not talking about a disease, he should have indicated that the labeling of the disease was general. By doing that, we were able to check whether the worker guessed the answers. The second question was asked to determine if the tweet was sarcastic or not.

|  |  |  |  |
| --- | --- | --- | --- |
| Fibromyalgia | Asthma | HIV | Annotations |
| 292 | 196 | 453 | Full consent |
| 4 | 26 | 27 | Consent |
| 0.424 | 0.563 | 0.223 | Fleiss' kappa |

Table 5: For the question 'Does the tweet talk about disease?', we checked the Fleiss' kappa between the workers.

Using the findings presented In Table 5, we checked how many tweets had full agreement amongst the workers. We wanted to check what the Fleiss' kappa values would be for such a simple question, although only a small portion of tweets for each disease did not have full consensus. The Fleiss' kappa values were barely sufficient; therefore it was necessary to let more workers answer the questions in order to get a better rating.

|  |  |  |  |
| --- | --- | --- | --- |
| Fibromyalgia | Asthma | HIV | Annotations |
| 291 | 188 | 422 | Yes -Full consent |
| 1 | 16 | 31 | No -Full consent |
| 4 | 10 | 7 | Yes - consent |
| 0 | 8 | 2 | No - consent |

Table 6: For the question 'Does the tweet talk about disease?', the results of the distribution of workers labeling.

Results Table 6 shows the workers' decisions. We saw, that even though there was full agreement in most cases, the Fleiss' kappa values were still relatively low. If we look at the Fleiss' kappa formula, we will understand that although there was a full consensus the entropy obtained for each disease is not good and therefore it effects the Fleiss' kappa. Therefore, for tagging tasks, Fleiss' kappa does not always represent the true quality of the task.

|  |  |  |  |
| --- | --- | --- | --- |
| Fibromyalgia | Asthma | HIV | Annotations |
| 249 | 159 | 293 | Full consent |
| 47 | 63 | 169 | Consent |
| 0.074 | 0.095 | 0.198 | Fleiss' kappa |

Table 7: For the question 'Is the tweet sarcastic?', we checked the Fleiss' kappa between the workers.

The results described in Table 7 show that indeed the task to determine whether the tweet is sarcastic or not was very difficult. We received a very low Fleiss' kappa because there was a lot of different possible answers. In Table 8 we presented the responses selected by the workers, that helped us understand why the Fleiss' kappa rating is poor.

|  |  |  |  |
| --- | --- | --- | --- |
| Fibromyalgia | Asthma | HIV | Annotations |
| 1 | 1 | 13 | Yes -Full consent |
| 249 | 158 | 280 | No -Full consent |
| 4 | 13 | 118 | Yes - consent |
| 42 | 50 | 51 | No - consent |

Table 8: For the question 'Is the tweet sarcastic?', the results of the distribution of workers labeling.

In table 8, we see the dispersion of the workers' responses. Unlike table 6, the answers are more scattered, so the Fleiss' kappa rating is very poor. Because there was no agreement amongst all the workers on the same answer or a limited range of answers, we received a very poor rating. Therefore, it is necessary to increase the number of workers in order to get a better rating and thus get better labeling.

Conclusions

After examining the results, we reached the following conclusions:

1. MTurk is a good system for collecting labels for classified training.
2. We received results in a reasonable amount of time - we did not see any difficulty in answering the questionnaires, although we began to have doubts about the clarity of the questionnaire.
3. There was a correlation between the results of the workers' labels and the results of the expert's labels.
4. Regarding the labeling of sarcasm, we have not always found a match between the experts and the workers - This can be explained by the fact that the subject of diseases was an issue, in some cases, of personal background and language proficiency.
5. There is a limitation on the number of characters (140) per tweet on Twitter, as a result a new language has 'developed', which contains abbreviations and hashtags to meet the characters limit, so even native English speakers will have difficulty understanding the tweets.
6. The workers' working time was very fast (72 seconds on average) and the quality of the work was very good. We did not notice a reduction in the quality of work if the worker solved more HITs.

In conclusion, MTurk is a very good system for tagging tasks. The system is very fast, versatile and easy to use. However, it is necessary to explain and simplify the HIT as much as possible in order to obtain good labels.

Acknowledgments

We thank to Dr. Minkov Einat for her help.

References

Rion Snow, Brendan O’Connor, Daniel Jurafsky2 and Andrew Y,ng 2008. *Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks,* Association for Computational Linguistics. [1]

Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. *Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?* Association for Psychological Science. [2]

Aniket Kittur, Ed H. Chi and Bongwon Suh. *Crowdsourcing User Studies With Mechanical Turk: Amazon.com’s Mechanical Turk*. [3]