CS 221 Project Proposal

Guoqin Ma, HAOTIAN LI

TOTAL POINTS

10 / 10

QUESTION 1

1 Proposal **10 / 10**

✓ **+ 10 pts Excellent**

+ 9 pts Good job

+ 8 pts Good job

+ 7 pts Meets requirements

CS 221 Project Proposal

Guoqin Ma, Haotian Li

Project name:

Naive Bayes Classifier and Neural Network in Tweets Classification under Disastrous Events

Baseline code and datasets: https://github.com/sebsk/AI_Project

1. Motivation

Social media is playing an increasingly important role in information collection and dissemination and it is changing our communication pattern significantly. Communication via social media platforms features large volume information, wide networks, quick responses, low costs, and intensive interactions. Especially during disastrous events such as flood, wildfire, earthquake, etc., posts posted by people involved in these disasters or bystanders sometimes could send out information much more early than traditional mass media. Thanks to these characteristics, since the rise of social media platforms such as twitter and Facebook, they have become an effective channel for authorities to collect information when these extreme events occurred. However, the quality of information could not be guaranteed due to the participation from vast number of users and its autonomous nature. A large fraction of the information collected from these platforms might be rumors, fake news, or totally irrelevant. As Stieglitz et al. (2018) mentioned in their article, "We are still in the early stages of developing appropriate methods and information systems to overcome the existing risks and challenges of using these platforms under disaster and emergency conditions." Therefore, in order to make better use of social media to facilitate decision making in natural disasters, it is imperative to build up an accurate and efficient text and image classifier to identify the relevance of collected data, such as tweets.

2. Task Definition

In this project, we will focus on text classifier.

When utilizing those data from social media platform, we may encounter some difficulties. For example, the amount of information will be extremely large, but not all of them are relevant and effective. This requires our classifiers being able to identify valuable information from unusable ones.

To handle the huge data volume, we need to make our algorithm efficient enough. Furthermore, since we are using the filtered information for surveying or analysis, it will be helpful if those "noise" could be minimized. Hence, an accurate classifier is wanted.

Therefore, the technical aspect is to improve the efficiency and accuracy of the classifier.

Input/output pairs

Input: tweets (text)

Output: relevance and **informativeness** to disastrous event

For example,

Input: "RT @BreakingNews: Magnitude 7.9 earthquake strikes off the coast of Costa Rica – USGS"

Output: Related and informative

3. Approach

Baseline

Logistic Regression with CountVectorizer preprocessing method is set as the baseline text classifier.

Proposed Methods

Considering its considerable available packages resources, Python is the programming language adopted in this project. The machine learning models, including logistic regression and naive Bayes classifier will be coded based on the machine learning toolbox named "sklearn" of Python. The deep learning network will be constructed with keras. As for text preprocessing, we will explore various techniques, including but not limited to regex, tfidfVectorizer (sklearn), nltk library, gensim library, etc. to compare their accuracy with CountVectorizer (sklearn).

In this project, we will do a large amount of work in text preprocessing, constructing the modified naive Bayes classifier and deep learning neural network and tuning all the parameters and structures of our model. Besides, we will also study the impact of training dataset and test dataset to answer questions related to generalization. 2 examples are "will a model trained by a flood event dataset work well on wildfire event?" and "will a model trained by an event in the United States work well on one that occurs in Australia?"

4. Evaluation

Metric

As stated in task definition section, our objective is to improve the efficiency and accuracy of the classifier. Therefore, one way for evaluation is to test the speed of our algorithm. Apart from this, the identification accuracy is also of vital importance. Hence, we may evaluate our algorithm by checking the precision and recall of the test results. A good classifier can accomplish the identification of a large amount of data in a short period of time, with high precision as well as recall.

Data

We will use 4 datasets, namely, 2012_Colorado_wildfires, 2013_Australia_bushfire, 2013_Colorado_floods, 2013_Queensland_floods, in our projects. These datasets contain tweets.

- **Sampling method**: by keyword filtering from tweets included in the 1% sample at the Internet Archive.
- **Labels**: ~28,000 tweets (about 1,000 in each collection) were labeled by crowdsource workers according to informativeness (informative or not informative), information types (e.g. caution and advice, infrastructure damage), and information sources (e.g. governments, NGOs).
- **Data format**: comma-separated values (.csv) files containing tweet-ids for the unlabeled tweets, plus the text of the tweets and labels for the labeled ones. Also includes a JSON file with metadata about the collection, including the keywords used to select tweets.

Qualitative Analysis

We could study about the impact of selection of datasets.

5. Plan

Team Roles

Guoqin Ma	Haotian Li
Literature Review	
Data Collection	Data Manipulation
Text Preprocessing	Data analysis
Naive Bayes classifier & Deep learning	

Timeline

Literature review and data collection	Done	Milestone: Project Progress Report	May 25
Milestone: Project Proposal	May 4	Deep learning	June 1
Data Manipulation	May 8	Milestone: Project Poster	June 5
Text Preprocessing	May 13	Data analysis and wrap-up	June 10
Naive Bayes classifier	May 20	Milestone: project final report	June 11

6. References

A. Olteanu, S. Vieweg, C. Castillo. 2015. What to Expect When the Unexpected Happens: Social Media Communications Across Crises. In Proceedings of the ACM 2015 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '15). ACM, Vancouver, BC, Canada.

Stieglitz, Stefan, Deborah Bunker, Milad Mirbabaie, and Christian Ehnis. 2018. "Sense-Making in Social Media during Extreme Events." Journal of Contingencies and Crisis Management 26(1):4–15. Retrieved April 17, 2018 (http://doi.wiley.com/10.1111/1468-5973.12193)

1 Proposal **10 / 10**

✓ **+ 10 pts** **Excellent**

+ **9 pts** Good job

+ **8 pts** Good job

+ **7 pts** Meets requirements