Language Classification

Team name: Cereal Killer

Paul Wu

#260727743

jiapeng.wu@mail.mcgill.ca

Andre Kaba

#260493293

andree.kaba@mail.mcgill.ca

Jingyun Liu

#260657058

jingyun.liu@mail.mcgill.ca

1. **INTRODUCTION**

We present a classification model to discriminate between 5 different languages. Short sentences or snapshots of conversations as samples from each language are given as training set, and we experiment with classifiers with different properties, such as logistic regression, naïve bayes classifier that draws linear boundary, and K-Nearest Neighbors (KNN), decision tree that draws nonlinear boundary, and random forest that considers multiple classifiers altogether. We compare different approaches of preprocessing and feature extraction of the raw text input, including manual removal of special characters and encoding lexical features. Results show that encoding based on unigram characters with weights on raw data with no removal of the special characters gives the best performance, independent of our choice of classifiers. Among all the classifiers we experiment with, XYZ classifier achieves the highest accuracy of XX% and an F1 score of XXX on validation set.

1. **RELATED WORK**

One of the main focuses of language identification task is discriminating similar languages. Zampieri et al. (2014) [1] concluded and compared various methods, where the most commonly used features are n-gram character or words features. More specifically, there have been approaches using bag-of-words for language identification (2013) [2] and n-gram language models combined with Part-Of-Speech distribution(2013)[3].

Another focus of language identification task is where the data is collected from and the task is applied to social media data, such as Twitter [4] and forums. Most commonly used feature extraction methods includes ad-hoc rank-order statistic over character n-grams proposed by Scheelen et al. (2003) [5]. In addition, Lui et al. [6] presented a multi-lingual detection model where features are constructed by character n-grams and selected using Information Gain(IG), reaching an F1 score of 0.9.

In this project, however, we are given very limited data in terms of text length and language variety compared to the previous works. Moreover, the text content is rather heterogeneous and contains tokens outside of the language. Thus we shall only experiment with the most common feature extraction methods, n-gram language model over the characters, after careful preprocessing.

1. **PROBLEM REPRESENTATION**
2. *Data Preprocess*

We observe that in the both training and testing data, there exists a mixture of corrupted and outliner characters that obviously do not belong to the language. Three approaches were compared: 1) remove outside of vocabulary for each language separately, in reference to the alphabet dictionaries in the language; 2) only remove emojis and urls while leave the others unchanged, since emoji and url are universal and clearly should not belong to any particular language; 3) do nothing other than converting everything into lower case, since the characters given in test case are all in lower cases. We concluded that method (3), doing minimal removal of the corrupted characters results in the best performance, thus is the final choice for predicting on test data.

1. *Feature Extraction.*

Previous works have shown that it n-gram language model with term frequency-inverse document frequency (TF-IDF) weighting is one of the most effective way of encoding inputs. We chose character instead of word representations since word representations would have a much higher dimension caused by word diversity across different languages, whereas character contains more compact information. Observing that the test data were single characters drawn from different language distributions instead of complete words or sentences, it makes more sense to encode the input at a character level using unigrams.

We observe that there are larger variety of characters in training set than test set. Therefore, we experimented on restricting the vocabulary to be the set of characters that appear in both training and testing set, in order to eliminate irrelevant and unnecessary noises for inference. In theory this operation reduces dimensionality of the feature space, and thus should improve the performance on some models such as decision tree, since the tree would have fewer split boundaries to consider and would need fewer splits. However, in practice we observed no significant improvement over the test set.

1. **ALGORITHMS IMPLEMENTATION**

Training data is split into 10 folds for cross-validation. We examine and collect accuracy and F1 score produced under different setups for comparison among classifiers. Since we notice that the vocabulary for test set is much smaller than the ones for training set, we suppose that some features are inherently more important than others, namely the ones that represent characters that appear in both training and testing. Thus we experimented with Principal Components Analysis (PCA) to reduce the dimensionality of the data by projecting the features onto an orthogonal space where variance is best explained.

We implemented the following classifiers:

1. *Linear classifiers*

**Naive Bayes**: Naive Bayes models have been widely used for clustering and classification. We assume that each feature in our data input follows a multinomial distribution, since each entry is a discreet real number representing the occurrence of a character. It is unreasonable to assume distributions such as Gaussian or Bernoulli, since the feature entries represents occurrence, and are not continuous nor binary.

The model estimate for each feature using Maximum Likelihood Estimator (MLE), and discriminates between classes by calculating for each label y. Because MLE is prone to overfit the observed distribution, we also experimented with different amount of smoothing, since Maximum a Posteriori (MAP) estimator results in a more generalized classifier.

Direct multiplication of fractions over times is numerically unstable and can cause underflow issues. In our implementation, we take natural log over all probabilities and do addition when calculating joint probability.

**Logistic Regression**:

1. *Nonlinear classifiers*

**KNN**: In vanilla KNN, inference time grows linearly with the number of training data point and dimension of features. Since the input matrix is sparse and lots of entries are 0, computation power were wasted on dot product operation. Thus we incorporated an Invert Index Based method (<https://arxiv.org/pdf/1011.2807.pdf>). The idea is to only do calculation on features where the point has a valid value and skip all the 0’s.  However due to high dimensionality of the input, this method still does not scale up. Thus although theoretically doable, we are unable to evaluate KNN efficiently with other models.

**Decision Tree**:

**Random Forest**:

1. **TESTING**

We conduct experiments on the performance of different classifier with various choices of preprocessing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Naive Bayes (laplace smoothing) | Logistic Regr (L2 reg) | Decision Tree | Random Forest |
| No preprocess, no tfidf | 0.855 | 0.859 | 0.855 | 0.911 |
| No preprocess, with tfidf | 0.717 | 0.877 | 0.85 | 0.912 |
| Remove emoji/url, with tfid |  |  |  |  |
| Remove all special | 0.818 |  |  |  |
| Vocab intersection |  |  |  |  |

Table X: Comparison across models with different preprocessing

We also examine the impact of PCA, by comparing performances of different classifiers trained on transformed data, keeping the setups the same. However, the performances show no significant improvement after the operation, and one reason might be that the input dimensionality is already rather small.

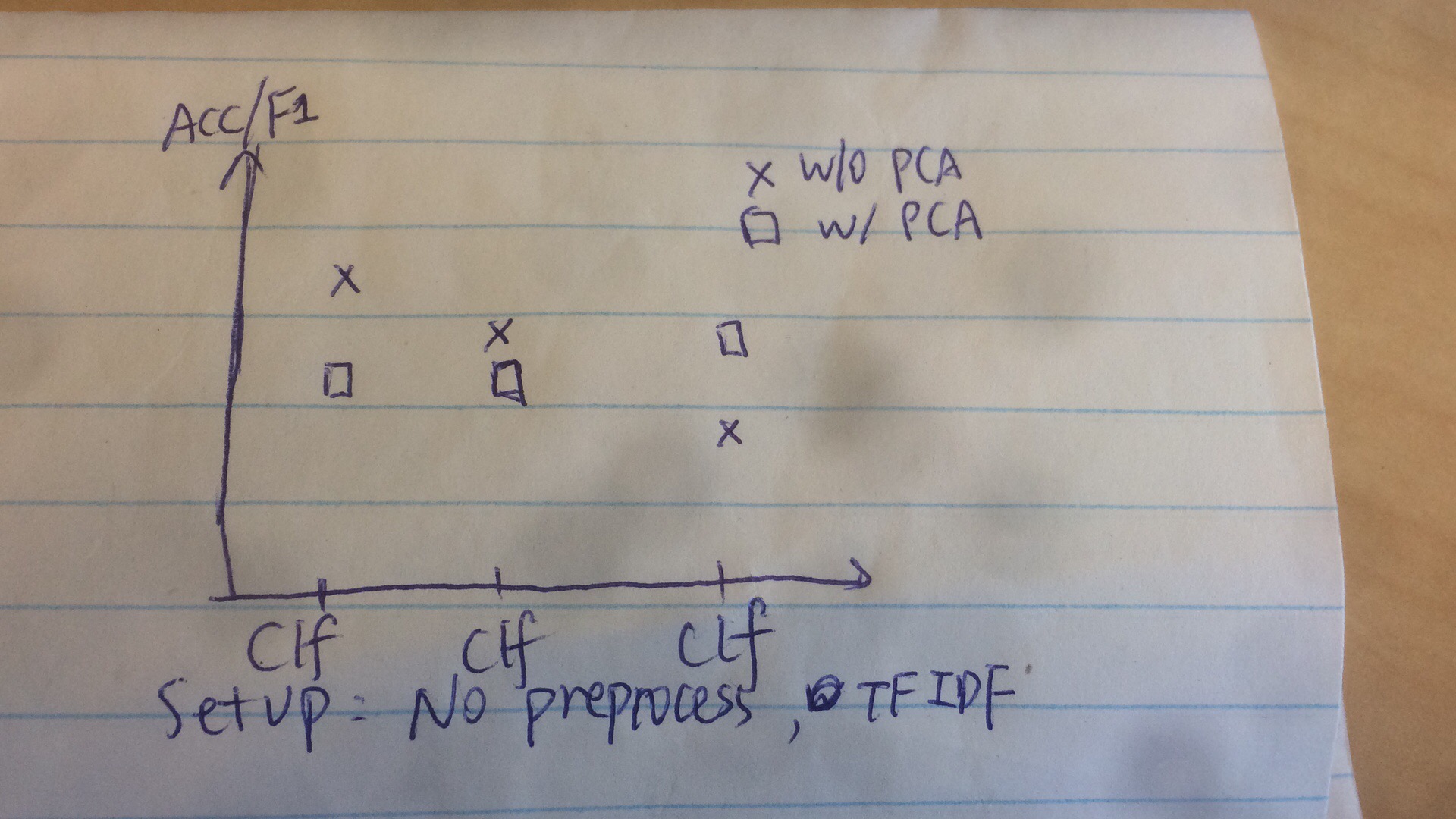


Figure X: Comparison of models with PCA

1. **DISCUSSION**

Unigram language model in general lacks robustness because it takes no consideration of any information given around target character, but it makes sense here since test set is composed of single characters drawn independently across language distributions. If the test data given are snapchats like phrases or sentences, we will have to consider more sophisticated language models for feature extraction, and maybe even classifiers that model the interaction between features.

1. **STATEMENT OF CONTRIBUTION**

Jingyun Liu:

* Implemented Naïve Bayes, KNN classifier
* Drafted most parts of the report

Jiapeng Wu:

* Data preprocessing and feature extraction.
* Implemented Logistic Regression
* Construct the structure and write the report

Andre Kaba:

* Implemented Random Forest, Decision Tree

We hereby state that all the work presented in this report is that of the authors

1. **REFERENCES**

[1] Zampieri, Marcos & Tan, Nikola & Tiedemann. (2014). A Report on the DSL Shared Task 2014. 10.3115/v1/W14-5307.

[2] M. Zampieri, "Using bag-of-words to distinguish similar languages: How efficient are they?," *2013 IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI)*, Budapest, 2013, pp. 37-41.

[3] Zampieri, Marcos & Gebrekidan Gebre, Binyam & Sascha, Diwersy. (2013). N-gram Language Models and POS Distribution for the Identification of Spanish Varieties.

[4] Lui, Marco & Baldwin, Timothy. (2014). Accurate Language Identification of Twitter Messages. 17-25. 10.3115/v1/W14-1303.

[5] Frank Scheelen, 2003. libtextcat. Software available at <http://software.wise-guys.nl/libtextcat/>.

[6] Lui, M & Lau, J.H. & Baldwin, T. (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*. 2. 27-40.

[7] Wang, J & Lin, L, Huang, T & Wang & He, T. (2010). Efficient K-Nearest Neighbor Join Algorithms for High Dimensional Sparse Data. *arXiv preprint arXiv:1011. 2807, 2016.*