CS 839 Spring 2019, Project Stage 1

PROJECT REPORT

TEAM MEMBERS

1. Karthik Chandrashekar - karthik.chandrashekar@wisc.edu
2. Kartik Sreenivasan - ksreenivasa2@wisc.edu
3. Niveditha Hariharan - nhariharan@wisc.edu

Entity: Person Names

For Example: Paula Radcliffe, Cristiano Ronaldo, Alex Ferguson

Dataset Source: http://mlg.ucd.ie/datasets/bbc.html

We have tagged the person names entities using three different tags <name>[data]</name>, <fname>[data]</fname>, <lname>[data]</lname>

All three labels are considered names. We just tagged them explicitly for ease of parsing and to collect more metadata.

For Example: <name>Christiano Ronaldo</name>, <fname>Wayne</fname>, </lname>Rooney</lname>

Document Mark-up Statistics

	SET I	SET J
No of Documents	200	100
Mark ups	3462	1567

Total Occurrences of Names in the Dataset : 5029

Pre-Processing Rule

Each data in training set where none of the words were capitalized,were filtered out. This was done in order reduce the number of negative examples in the training data as suggested in the lecture.

Features Used

1. Feature_possessive_form - To check if the last word in the string ends with 's or '
2. Feature_num_capitals - The number of capital letters in the string
3. Feature_num_vowels - The number of vowels in the string
4. Feature_num_consonants - The number of consonants in the string
5. Feature_num_characters - The number of characters in the string
6. Feature_ascii_sum - The sum of ASCII values of the string
7. Feature_number_present - Checks if the string has a digit
8. * Feature_is_noun - If majority of the words in this string are tagged propernoun then the string is considered noun.
9. Feature_is_day_month - Checks if the string contains a day or a month
10. Feature_contains_stopwords - Checks if one of the words in the string contains stopwords
11. Feature_contains_city - Checks if one of the words in the string is a city
12. Feature_begins_sentence - Checks if the first word occurs first in the sentence
13. Feature_contains_sports - To identify common sports terms
14. * Feature_is_noun_v2 - Checks if a string is a noun using the context based dictionary
15. * Feature_is_verb_v2 - Checks if a string is a verb using the context based dictionary
16. Feature_contains_Countries - Checks if one of the words in the string is a Country

* We used nltk pos tagger to tag parts of sentence as proper noun,verb etc from which we derived the above * features.

Cross Validation Results on Set I

We explored various Machine Learning models for the given dataset. Here is an elaborate list of the models that we used and their Precision , Recall and F1 values when we did a 3 fold Cross Validation on the training data.

Machine Learning Model	P	R	F1
DecisionTreeClassifier(Entropy)	0.8389	0.8409	0.8399
DecisionTreeClassifier(gini)	0.8467	0.8433	0.8450

Logistic Regression	0.7286	0.4760	0.5758
RandomForestClassifier	0.8557	0.8538	0.8548
GaussianNB	0.3847	0.9849	0.5533
SupportVectorMachine_SVC	0.8058	0.7684	0.7867

Since Random Forest gave the best performance, so we chose **RandomForestClassifier** for the next step.

Before Post processing on Set J

Here is the performance of Random Forest before the rule based post processing step

P	R	F1
0.8845	0.8743	0.8794

After Post processing on Set J

The rule-based post processing was done in order to eliminate false positives. The post processing rules can be found in football.csv file of the code directory. These are mostly football club names (Example: Everton, Manchester United, Tottenham etc.) which the classifier identified as positive and some football related terminologies (Example: Coach, Manager etc.)

P	R	F1
0.9328	0.8330	0.8801

Final Results

These results Precision = 93.28% and Recall = 83.30% meet the expected Project requirements of at least 90% Precision and 60% Recall