

# CIS530 FINAL REPORT

STUART WAGNER & MICHAEL WOODS

## CONTENTS

1	Introduction	1
2	Methods and Resources	2
2.1	Data . . . . .	2
2.2	Tools and Software Packages . . . . .	2
2.3	Features . . . . .	3
2.4	Models . . . . .	5
3	Final system	5
3.1	Team . . . . .	5
3.2	System Design . . . . .	5
3.3	Models . . . . .	5
3.4	Features . . . . .	6
3.5	Pipeline . . . . .	7
3.6	Testing . . . . .	7
4	Outcomes	7
4.1	Submission results . . . . .	8
5	Discussion of Results and Conclusion	8
5.1	Analysis of feature performance . . . . .	8
5.2	Keys to success . . . . .	9

## 1 INTRODUCTION

Text classification on the basis of high or low information density has until recently <sup>1</sup> been a relatively overlooked task in the field of Computational Linguistics. Building off the contributions of recent research, we seek to improve accuracy on correctly classifying documents with high information density using the tools and techniques of machine learning.

Information density classification has many practical applications, most of which relate to the task of automatic summarization. Text summarization has relied partially on the use of KL divergence—the measurement of a difference between one probability distribution and another—in order to determine the quality of a generated summary. However, this measure fails to account for information density, a rather abstract metric of the “informativeness” of a given piece of writing.

Nenkova and Yang used a number of features in developing their model. We used many similar features, however as students new to NLP, we instead

---

<sup>1</sup> Nenkova & Yang “Detecting Information-Dense Texts in Multiple News Domains” 2014  
<http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8430/8622>

thought that using a variety of learners could potentially provide better results. We therefore approached the problem as a machine learning problem that leveraged our limited knowledge of NLP for the purposes of feature engineering. Could we learn which features were more correlated with information density? Could we leverage a variety of models, using ensemble learning, to improve prediction accuracy? The answer to those questions, generally speaking, is yes.

## 2 METHODS AND RESOURCES

We decided to frame the problem of classifying a given article as either information dense or sparse as primarily a machine learning problem. As with most problems that can be addressed with the methodologies of machine learning, having some knowledge of the domain—in this case text classification—can provide useful insights when selecting and extracting features from raw data. With the approaches we learned in CIS530 regarding the analysis of the structural components of text, along with our own intuitions and experimental insights, we outline the various features, tooling, and resources we employed for the project below.

### 2.1 Data

Apart from the 2,282 training and 250 test leads we were provided for the project, we made use of the following third party data resources:

#### **MRC PSYCHOLINGUISTIC DATABASE**

[http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa\\_mrc.htm](http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm)

The MRC Psycholinguistic Database is a machine readable dictionary containing a number of linguistic and psycholinguistic attributes for each word<sup>2</sup>. Specifically, we primarily made use of the 4,923 word base lexicon included in the MRC distribution.

#### **LINGUISTIC INQUIRY AND WORD COUNT (LIWC)**

<http://www.liwc.net>

The Linguistic Inquiry and Word Count (LIWC) is a text analysis software program designed to calculate the degree to which people use different categories of words across a wide array of texts, including emails, speeches, poems, or transcribed daily speech.”<sup>3</sup>

### 2.2 Tools and Software Packages

#### **STANFORD CORENLP**

<http://nlp.stanford.edu/software/corenlp.shtml>

A suite of text processing tools capable of part-of-speech (POS) tagging, named entity recognition (NER), tokenization, lemmatization, sentiment analysis, as well as dependency relation and syntactic parse extraction.

---

<sup>2</sup> <http://www.psych.rl.ac.uk/>

<sup>3</sup> <http://www.liwc.net/>

**SCIKIT-LEARN**

`scikit-learn`<http://scikit-learn.org/>

A popular Python machine learning framework with implementations for a large selection of popular classifiers.

**NUMPY**

<http://www.numpy.org>

A scientific computing package for Python providing a native n-dimensional array datatype, as well as various linear algebra, statistical, and general mathematical functions suitable for “number crunching.”

**NLTK**

<http://www.nltk.org/>

A Python library for working with natural language that includes a number of convenient text processing tools and built in corpora for multiple languages.

### 2.3 Features

Building off the work of Nenkova and Yang, we leveraged a number of similar features in our approach to the problem of textual information density classification.

#### (1) CORENLP WORD COUNT—BINARY BAG-OF-WORDS

*9799 dimensions.* All unique tokens occurring between `<word>``</word>` tags in the generated XML obtained from running all training files through the CoreNLP tool were collected into a list and converted to lowercase. Any tokens consisting of stop words, as defined by NLTK’s English stop word list or punctuation characters were removed. The list was then sorted alphabetically in ascending order and each word was assigned an index. To construct the feature vector for a given lead, the text of the lead was processed in a manner similar to that of the XML output and a vector the size of the bag-of-words model was initialized with zeros. If the lead contained the *i*-th word in the bag-of-words model, the *i*-th position in the feature vector was set to 1.

#### (2) CORENLP SENTENCE INFORMATION

*7 dimensions.* To construct the feature vector, each processed lead had a number of pieces of information extracted from it directly or from the XML file generated from it using CoreNLP. These quantities define the features indicated below along with our prediction on its impact on information density (+/−):

- The number of sentences per lead. (−)
- The number of non-unique tokens per lead. (−)
- The number of named entities recognized in the lead. (+)
- The number of nouns in the lead (tokens tagged as NN, NNS, NNP, or NNPS) (+)
- The number of words with over six letters. (+)
- The number of quotation marks appearing in the lead. (+)

- The aggregate absolute sentiment score of the lead. Each sentence was assigned a numeric score according to the following rubric: “very negative”: 2, “negative”: 1, “neutral”: 0, “positive”: 1, and “very positive”: 2. The scores were then summed and the absolute value was taken. (—)

### (3) LIWC

*78 dimensions.* Our hypothesis was that words indicating sentiment and emotions might indicate sentences that were less information dense. We thought these might be positively correlated with information density: present tense, numbers, impersonal pronouns, and causation. We also believed multiple attributes may be negatively correlated, including: 1st person usage, auxiliary verbs and past tense, quantifiers, and positive and negative emotion. For a listing of the data points in the LIWC, refer to the table at <http://www.liwc.net/comparedicts.php>.

### (4) CORENLP PRODUCTION RULE COUNT—BINARY BAG-OF-WORDS

*5654 dimensions.* The syntactic parse generated by CoreNLP for each sentence from a given lead was extracted from the corresponding <parse></parse> tags in the corresponding XML file. A parse tree was constructed in the same manner as described in Homework 3, where only non-terminal nodes were considered. To construct the master list of production rules, the outlined process was repeated for every lead file in the training set and every unique production rule was collected into a list. Each rule was then assigned an index, 0 to  $N - 1$ , where  $N$  is the number of unique production rules. To construct a feature vector for a given lead, a  $1 \times N$  vector of zeros was created and every production rule occurring in the sentences of the lead was extracted. For every rule in the master list, if the  $i$ -th rule occurred in the rules extracted from the lead, the  $i$ -th index in the feature vector was set to 1. The belief was that perhaps certain production rules may be related with information density.

### (5) MRC DICTIONARY—BINARY BAG-OF-WORDS

*4923 dimensions.* The MRC Psycholinguistic database word dictionary consisting of 4,923 words was used to construct a binary bag-of-words model in a similar manner to the CoreNLP binary bag-of-words model. We also attempted counts normalized by length of lead, however there did not appear to be a significant difference in accuracy. The text of each lead was tokenized, converted to lowercase and stripped of stop words. The benefit of using this approach is that the dictionary is independent of the task and training data (Nenkova and Yang).

### (6) DEPENDENCY RELATIONS—BINARY BAG-OF-WORDS

*5000 dimensions.* For each lead file in the training set, the <dep></dep> nodes were extracted from the corresponding CoreNLP XML files and converted to a 3-tuple consisting of the dependency relation type, the governor word, and the dependent word. Each tuple instance was counted, and the top 5000 dependency tuples were used to build a bag-of-words vector in much the same manner as features (1), (4) and (5). Constructing the feature vector for a given lead proceeded in much the same manner as (5) as well. Like (4), we believed that certain dependency relations are significantly more prevalent in information dense articles than sparse ones.

## 2.4 Models

Regarding the actual task of classification, we constructed several classification models using the features described above. The model configurations are as follows:

- Logistic regression with a L2 penalty.
- Linear SVM with L2 loss and penalty functions.
- Bernoulli naive bayes with default parameters.
- Ensemble method with the three classifiers described above. Each classifier makes a prediction for a given observation and a majority vote is taken to determine the final predicted label.

# 3 FINAL SYSTEM

## 3.1 Team

We are team “W<sup>2</sup>,” or as it appears on the leaderboard in ASCII, W^2.

## 3.2 System Design

From the very onset of the project, we exercised great discipline in producing what we believe to be a clean, extensible system for making predictions about text. The software could easily have been treated as disposable—and designed as such—since the predictions produced by the system are ultimately the real value and not the software itself. Our rationale for putting a large, up-front effort into properly designing a data pipeline was to introduce a lower mental overhead cost when attempting to quickly evaluate the effectiveness of various featuresets and models as the project progressed. The effort paid off many times over, as creating new models and features was the same as defining a new Python module. The model or feature could then later be referred to by name and manipulated through set of generic functions.

## 3.3 Models

All models are implemented as regular Python modules in the `project.models` package in the project code directory. Additionally, all model modules are expected to define the functions described below, as this provides a common interface to interact with the model implementation:

```
def preprocess(model_name, train_files, test_files)
def train(model_name, train_files, labels, *args, **kwargs)
def predict(model_name, model_obj, test_files, *args, **kwargs)
```

### 3.3.1 Lookup

A module can then easily be looked up by name like so:

```
linear_model = get_model('linear_model')
```

which is defined in the following simple function which makes use of the `importlib` from the Python standard library:

```
def get_model(name):  
    return importlib.import_module('project.models.'+name)
```

### 3.4 Features

Much like models, features are also implemented as Python modules, as this allows for quick prototyping and iterative refinement of new feature types.

In a similar manner to models described above, features (or more accurately feature-izers) can easily be resolved with the following function:

```
def get_feature(name):  
    return importlib.import_module('project.features.'+name)
```

Features can also expose the following functions:

```
def build(feature_name, *args, **kwargs)  
def featurize(feature_name, feature_obj=None, *args, **kwargs)
```

where `build` performs the initial setup work needed to featurize an input, and `featurize` performs the actual conversion of an input into a feature vector.

### 3.5 Pipeline

The conceptual classification pipeline is outlined below in pseudocode, which describes the general flow of information in the system, initially from raw input text to final predicted labels:

```

TRAIN_FILES = framework.get_train_files()

if in-submission-mode:
    TEST_FILES = framework.get_test_files()
else:
    TEST_FILES = TRAIN_FILES

# if the model defines an optional preprocess() function,
# call it:
if is_preprocess_defined('foo')
    optional = preprocess('foo', TRAIN_FILES, TEST_FILES)
else:
    optional = []

model_obj = train('foo', TRAIN_FILES, TRAIN_LABELS, *optional)

return predict('foo', model_obj, TEST_FILES, *optional)

```

### 3.6 Testing

Finally, the above pipeline is invoked from within the context of a testing framework we designed. The framework itself provides the input files to the prediction pipeline, which may vary depending on whether an actual prediction submission is being generated. If a leaderboard submission is to be generated, then actual test data is supplied to the pipeline by the framework, otherwise some holdout from the training data is used as test data. After the predictions have been made, the framework provides basic reporting information regarding the accuracy, precision, and F-score of the results.

## 4 OUTCOMES

All experiments were conducted using the `scikit-learn` `cross_validation.train_test_split`, function with 10% of the training data held out as the test set.

All submission configurations and related results are summarized in the table below. All feature sets are referenced by number as given in section 2.3.

#### 4.1 Submission results

Attempt	Accuracy	Model	Feature sets
1	75.6%	Logistic regression	1, 2, 3
2	76.0%	Logistic regression /w PCA(n=500)+gridsearch	2, 3, 4, 5
3	78.4%	Ensemble (LR+SVM+NB) (tainted w/ class weights)	2, 3, 4, 5
4	78.4%	Ensemble (LR+SVM+NB) (tainted w/ class weights)	1, 2, 3, 4, 5
5	79.6%	Ensemble (LR+SVM+NB) (tainted w/ class weights)	2, 3, 4, 5, 6
6	79.2%	Ensemble (LR+SVM+NB) amended w/o class weights	2, 3, 4, 5, 6

## 5 DISCUSSION OF RESULTS AND CONCLUSION

Overall, we are quite pleased with our results. Each test submission resulted in a general trend of higher predictive accuracy. We believe this is due to the design of our system which allowed for quick a test-prototype-report cycle of various model and feature configurations, which in turn encouraged experimentation.

#### 5.1 Analysis of feature performance

All told, we developed seven different feature sets:

1. [CoreNLP word count;binary bag-of-words]
2. [CoreNLP word count;TDIDF normalized bag-of-words]
3. [CoreNLP sentence information]
4. [LIWC]
5. [CoreNLP production rule count;binary bag-of-words]
6. [MRC dictionary;binary bag-of-words]
7. [Dependency relations;binary bag-of-words]

##### 5.1.1 *The good*

Of the above features, [CoreNLP sentence information], [LIWC], [CoreNLP production rule count;binary bag-of-words], [MRC dictionary;binary bag-of-words], and [Dependency relations:binary bag-of-words] performed very well, usually achieving a prediction accuracy in the 80 – 85% range when each featureset was used to make predictions in isolation with a logistic regression classifier. When used together in the ensemble model, the predictive accuracy spanned 80 – 89% when running against a cross-validated test set.

##### 5.1.2 *The not-so-good*

While the performance of the featureset [CoreNLP word count;binary bag-of-words] was not terrible, it was not great either. When used as a feature set in isolation with a logistic regression classifier, prediction accuracies generally failed to break 75%. This could be due to the fact that a binary bag-of-words model regarding the distribution of information dense/sparse articles is simply just not that informative, meaning the words that appeared in information dense and information sparse articles were not different enough to act as a good discriminative metric.



### 5.1.3 *The bad*

The [CoreNLP word count;TDIDF normalized bag-of-words] featureset consistently performed the worst, routinely failing to break 70% during testing and achieving an average prediction accuracy in the 60 – 65% range. The poor performance of this featureset was probably due to similar reasons as that of the [CoreNLP word count;binary bag-of-words] featureset, combined with a potential error in the implementation.

## 5.2 Keys to success

The ensemble model, we believe, allowed our system to outperform most other teams, even when using similar feature sets. The intuition behind the success of such an approach is straightforward: when in doubt, seek multiple opinions in order to come to an answer. The trade-offs when using an ensemble approach to prediction are mostly related to time—instead of training a single classifier, three, five, etc. classifiers need to be trained instead. Also, it is necessary to use an odd number of classifiers that achieve roughly the same level of predictive accuracy when used with the same feature set.