Carmen Easterwood

Project Milestone

I. Abstract

In this paper, I develop a topic classification model for news articles. I test multiple different model inputs, but find that feeding the full article text into the model still generates the best results. The highest performing model (so far) is a logistic regression model with accuracy of 77%, and F1 scores to be calculated for the final paper.

II. Introduction (motivation for your work)

Classifying news articles by topic is an important form of content analysis. It lets us know what journalists are writing about, whether that is changing over time, and the relative important of different topics. Tagging articles by topic also makes article searches more efficient for news consumers and researchers. However, manual topic-coding is quite time-consuming, so automating this process can save time and money for news organizations. This paper will develop a supervised topic classification model using New York Times news articles.

GitHub repo: https://github.com/carmen16/w266-final-project

III. Background

There is a large literature looking at various aspects of document classification, but only a few have focused on evaluating newspaper articles. Martin and Johnson (2015) used a dataset from the San Jose Mercury News and found that using a nouns-only or lemmatized version of the articles improved topic classification. However, theirs was an unsupervised model, so they evaluated their model using techniques that are not applicable in a supervised setting.

Meanwhile, Wermter and Hung (2002) created a semi-supervised self-organizing memory (SOM) model for topic generation of Reuters news articles. They used WordNet relationships to assist their model and thus limited their model input to only the nouns and verbs found in WordNet. This allowed them to achieve accuracies over 95%, but generalizability may be limited due to this use of WordNet.

Other recent papers have looked at supervised topic classification for non-news documents. Karan et al (2016) looks at classification of Croatian political texts) using word2vec, logistic regression, Gaussian NB, and gradient-boosted trees, as well as a couple of postprocessing rules, to reach an F1 score of 77% for major topics. Glavaš et al (2017) looks at SVM and CNN models for multiple languages, and for monolingual English models finds that CNNs perform best with an accuracy of 57%. In this paper, I will test whether some of the ideas and techniques used for classification of non-news documents can be successfully extended to the news article setting.

IV. Methods

Data Set

I am using the New York Times Annotated Corpus, which contains 1.8 million articles from 1987-2007, labeled with topics, subtopics, and newspaper sections. These articles come in individual XML files, and I have selected a random sample of 10,000 articles to parse and use as data for the model. The articles are randomly split so that approximately 75% are training data, 5% dev data, and 20% test data.

Model Input

I use the full text of each article as input to the models, but I also test four other model inputs based on my review of the literature. Based on Martin and Johnson (2015), I have created a nouns-only version and lemmatized version of each news article, and I will be adding headlines per Wermter and Hung (2002). I have also chosen to test the lead paragraph of each article, since news articles typically use the first paragraph to summarize all of the most important information in an article.

The table below shows the vocabulary size and the average number of unique words per article for each type of model input for a random sample of 10,000 articles. For each set of training data, I TF-IDF vectorize it before feeding it into models.

Model Input	Vocab Size	% of Full Vocab	Avg. Unique Words per Article
Full Text	85,031		260
Lead Paragraph	39,378	46%	61
Headlines	TBD	TBD	TBD
Nouns Only	68,324*	80%*	120*
Lemmatized	79,931	94%	245

** Will revisit this. Martin & Johnson (2015) found 36% of the San Jose Mercury News vocab was nouns.*

Model Output

The NYT dataset has multiple ways of classifying articles, and the table below shows examples of the different ways eight articles were classified. I have chosen to use the ***desk*** column as my output variable because (1) it has the lowest percentage of null values, and (2) it never assigns multiple descriptions to the same article.

Desk	General Descriptor	Online Sections	Taxonomic Classifier
Foreign Desk	['Immigration and Refugees', 'Jews', 'Music', 'Religion and Churches']	World	['Top/News', 'Top/News/World/Countries and Territories/Austria', 'Top/News/World/Europe', …
Metropolitan Desk	['Murders and Attempted Murders', 'Basketball', 'Culture', 'Children and Youth']	New York and Region	['Top/News/U.S./U.S. States, Territories and Possessions/New Yorks', 'Top/News/U.S./Mid-Atlantic', 'Top/News/Sports/Pro Basketball', …
Book Review Desk	['Books and Literature']	Arts; Books	['Top/Features/Books/Book Reviews', 'Top/Feature/Arts', 'Top/Features/Books']
Editorial Desk	NaN	Opinion	['Top/Opinion/Opinion/Letters', 'Top/Opinion', 'Top/Opinion/Opinion']
National Desk	NaN	U.S.	['Top/New/U.S.', 'Top/News']
Metropolitan Desk	NaN	New York and Region	['Top/News/New York and Region']

Classified	NaN	Paid Death Notices	['Top/Classifieds/Paid Death Notices']
Leisure/Weekend Desk	NaN	Arts	['Top/Features/Arts']

I have done some clean-up on the output variable to deal with misspellings and other errors (e.g. "Classifeds", "Classifieds;", and "Classified Desk" are all grouped together under "classifieds").

Models

I use multinomial naïve Bayes and multi-class logistic regression (one vs. rest) as simple baseline models. I will also be testing more complicated models based on my review of some relevant literature (SVM, CNN, and gradient boosted trees).

V. Results and Discussion

For the two baseline models, using the full text results in the highest accuracy. This differs from Martin and Johnson (2015), who found that lemmas and nouns perform better; however, their results are for unsupervised topic classification, which may benefit from having a more refined set of words to evaluate since the model also needs to generate the topic names themselves. A supervised model, meanwhile, appears to benefit from having as many words in the vocabulary as possible.

Model Type	**Input**	**Parameters**	**Accuracy**
Multinomial NB	Full Text	$\alpha = 2^{-7.5} = 0.00552$	0.707
Logistic Regression (one vs. rest)	Full Text	penalty = L2, C = 100	0.768
SVM			
CNN			
Gradient Boosted Trees			

I will also be looking at the best model for each input type and checking for any interesting results there.

	Best Model		
Model Input	**Type**	**Parameters**	**Accuracy**
Full Text			
Lead Paragraph			
Headlines			
Nouns Only			
Lemmatized			

VI. Next Steps (section for work you plan to do before submitting the final version)

My next steps are:

- Implement neural models (SVM, CNN) and gradient-boosted trees
- Error analysis by topic and by individual articles

- Possible addition of post-processing rules (see Karan et al (2016))

Finally, an interesting future project would be to run an unsupervised topic classification algorithm on this corpus (Martin and Johnson (2015) use the Latent Dirichlet Allocation algorithm) and compare those results to the supervised learning results.

VII. References

Fiona Martin and Mark Johnson. 2015. More Efficient Topic Modelling Through a Noun Only Approach. Proceedings of Australasian Language Technology Association Workshop, pages 111-115.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-Lingual Classification of Topics in Political Texts. Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, pages 42-46.

Mladen Karan, Jan Šnajder, Daniela Širinić, and Goran Glavaš. 2016. Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Texts. Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 12-21.

Stefan Wermter and Chihli Hung. 2002. Selforganizing Classification on the Reuters news corpus. Proceedings of the 19th international conference on Computational Linguistics – Volume 1.