News Article Text Classification and Summary for Authors and Topics

Aviel J. Stein[1], Janith Weerasinghe[2], Spiros Mancoridis[1], Rachel Greenstadt[2]

[1]College of Computing and Informatics, Drexel University, Philadelphia, Pennsylvania, USA
Ajs568@drexel.edu, Mancors@drexel.edu
[2]Tandon School of Engineering, New York University, New York, USA
Janith@nyu.edu, Greenstadt@nyu.edu

Abstract

News articles are important for providing timely, historic information. However, the Internet is replete with text that may contain irrelevant or unhelpful information, therefore means of processing it and distilling content is important and useful to human readers as well as information extracting tools. Some common questions we may want to answer are "what is this article about?" and "who wrote it?". In this work we compare machine learning models for evaluating two common NLP tasks, topic and authorship attribution, on the 2017 Vox Media dataset. Additionally, we use the models to classify on a subsection, about ~20%, of the original text which show to be better for classification than the provided blurbs. Because of the large number of topics, we take into account topic overlap and address it via top-n accuracy and hierarchical groupings of topics. We also consider edge cases in authorship by classifying on inter-topic and intra-topic author distributions. Our results show that both topics and authors readily identifiable consistently perform best when using neural networks rather than support vector, random forests, or naive Bayes classifiers, although the latter methods perform acceptably.

Key Words

Natural Language Processing, Topic Classification, Author Attribution, Summarization, Machine Learning

1. Introduction

The Internet is full of information, and a large part of it is text and images. Images are fast for humans to process but text takes more time. Natural Language Processing (NLP) techniques use statistical and computation driven methods to analyze large bodies of text. One of the most common forms of text online is a news article. In Section 2, we discuss related work in NLP. Two common tasks for NLP scientists is either authorship or topic classification. Authorship classification can be useful for plagiarism or detecting fake accounts and topic classification can be helpful for sorting or searching a dataset. The 2017 Vox Media is an understudied dataset that has advantages over other contemporary news article datasets in terms of the number of articles as well as labeled topics and authors. Most studies only explore one of these tasks, so one advantage of this work is that we explore both side-by-side in the same context, and, thus, showing that they are comparable techniques. Another item we explore is how extractive summaries of text can help distill important information from larger texts for either human or model consumption. These NLP techniques are helpful for many academic and industrial applications as off-the-shelf, open-source tools have become more reliable and accessible. Because contexts may differ, it is important to have baselines and reusable datasets to compare results or build models for transfer learning. One such dataset is the "20 newsgroup text dataset", which contains around 18,000 articles on 20 topics and does not include author labels. By contrast, Vox Media published a dataset that includes approximately 23,000 articles covering 186 topics and 817 authors. The Vox Media dataset [1] was published in 2017 and has received surprising little attention from the NLP community.

In Section 3, we discuss what methods we use to extract features and classify text. Text classification generally relies on machine learning to provide high accuracy results when applied to large data sources. For our two classification tasks, authorship attribution and topic classification, we extracted several types of features such as word n-gram, term frequency inverse document frequency (TFIDF), and part of speech (PoS) features but found that n-gram word count resulted in the best performance. We perform classification with various common machine learning models (see Section 3.3). Text summarization is performed by distilling the most important pieces of the text to a suitable degree of the original text. We used word frequency as the words score in each sentence and found the sentence score by averaging the score of all the words in a sentence, baring stop words. We constructed two types of dataset for topic and author, the dense dataset contained 10 classes each with 300 samples and the sparse dataset contained 50 classes each with 50 samples.

In Section 4, we perform the experiments demonstrating NLP efficacy for Vox articles. After performing the classifications, we inspected our models by performing confusion matrix and feature analysis, to understand how the classification may be affected by a confluence of signals. Previous work [2] on the Vox Media dataset explored the use of unsupervised learning to identify topics and categories of articles. This is a good approach, since several of the topics are closely related (*e.g.,* politics *vs.* politics and policy). We also used some unsupervised approaches to explore what kind of commonalities the texts exhibited regardless of their labeled class. To account for this, we used a top-n accuracy and 2-layer hierarchical approach. We categorize similar topics as into groups and first classify on the main topic, then categorize the sub-topics within each category. To account for authorship possible edge cases, such as all authors writing about the same topic or each author never writing about the same topic more than once, we also constructed inter-topic and intra-topic datasets and found that in both cases the authorship signal is still strong, sometimes stronger than the topic signal. Generally, though, authors tend to write about the same topics as they have in the past. For the 10-class dataset we attained 74% accuracy topic attribution and 86% accuracy author attribution. Using the same methods to extract features from the summaries of the 10-class dataset, we obtained 60% and 53% accuracy for topics and authors respectively. Summaries retained the authorship signal because they consist of a subsection of sentences from the original text. these summaries contained valuable information for machine learning models than the original summary, or "blurb", provided by the dataset. Correcting for topic overlap, with top-n and hierarchical models we can attain topic attribution between 83%-87%. We also considered inter-topic and intra-topic authorship attribution and found that with similar conditions to the dense dataset, in this case 8 authors with 300 samples each, authorship can be attributed with up to 92% accuracy in inter-topic. Intra-topic is a little harder, with only 50 samples each and 8 authors, it scores 68% accuracy. Finally, in Section 5, we consider the limitations of these approaches, discuss the implications of our work, and suggest ways that future research can improve upon them.

2. Related Work

2.1 Topic Classification

There are many aspects of text that can be attributed beyond topics as well such as classifying news based on bias [3] (see Figure 1) and credibility [4] as well as detect fake news [5]. For example, one approach classifies news articles based on their source and attributed to Fox, Vox, or PBS with at best 94% accuracy, but is it because of the text's style or the content signal?[6] The approach used by Yirey et al.[7] focuses on distinguishing between articles on Finance, Stocks, Education, and Environment and scores around and with a similar number of articles per topic. However, one drawback was that the dataset had to be well balanced. Another use of topic analysis is tracking topics that a user may be interested in and can help suggest future articles for the user to read.[8] This of process has to do two things, 1) track articles a user

reads, and 2) identify the topics articles. Topics of past articles will likely be similar to topics in future articles. An open question is whether a user likes articles written by certain sources or authors.

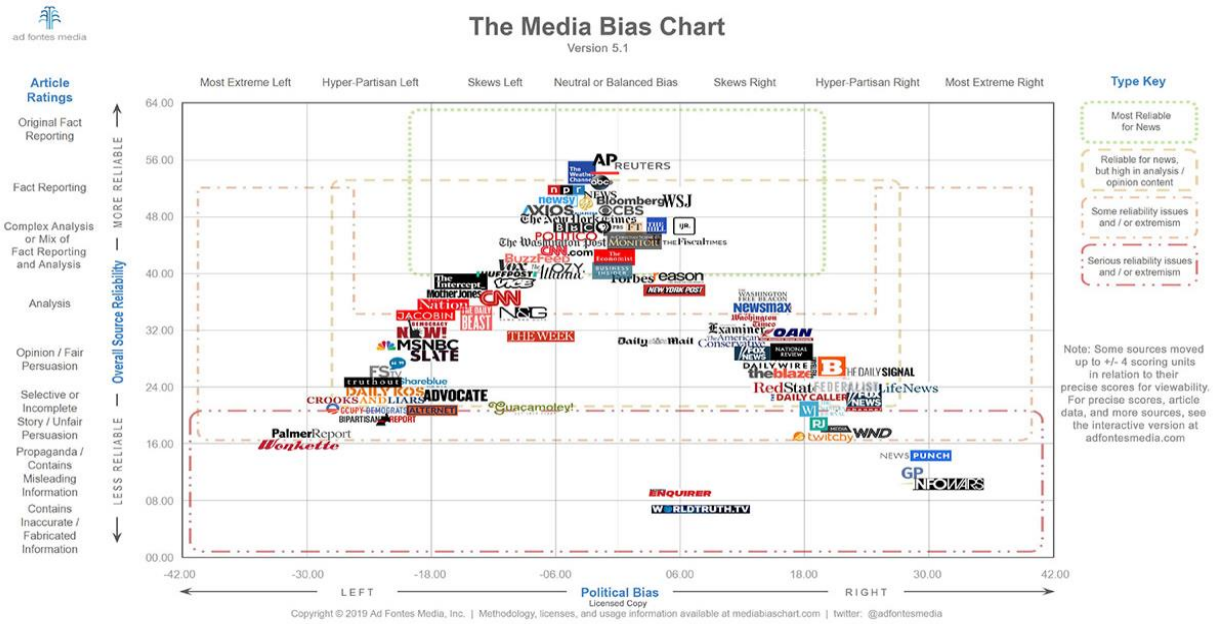

Figure 1. News organizations by political bias and overall reliability. Vox articles skew to the left of the political spectrum and are generally a "complex analysis or mix of fact reporting and analysis"

2.2 Authorship Classification

A related problem involves attributing individual authorship to documents. While this is not strictly an NLP task, as it has also been applied to other things where authorship is relevant such as art[9], music[10], source code[11], *etc.*, it is most prevalent with text. Authorship classification can be used to determine if someone plagiarized or helped preserve the anonymity of the author. Researchers performed deep learning authorship attribution on a dataset of 10 authors and lengthy articles achieving 95% accuracy, and, on shorter articles, 77% accuracy.[12] However, in the case of a news organization, they may have tens or hundreds of authors, so it may not be as robust in those circumstances. In less edited and smaller text portions, whiteprints scored 95% accuracy on eBay comments.[13] The level of professional editing could influence how much style is present. Another difference between this and other text corpora, is that the authors of articles likely pass their work through editors, and also likely have a style guild. This may create organizational signal or decrease authorship signal.

2.3 Text Summarization

Text summarization is the processes of generating a condensed document that retains the meaning and important information from the original text source. We generate a summary by using a small fraction of the text from the original. The two main ways to generate summaries are extractive and abstractive. Abstractive is harder and requires sophisticated learning and NLP approaches to produce novel phrasing. We choose to go with extractive because they filter for the most important sentence and are easy and flexible to construct. There are also two different kinds of summaries, inductive and informative. Inductive tend to be very short (~5% of original text) and informative are longer (~20% of original text). [14,15,16] In their survey of existing summarizing methods, they compare these methods but take the categories for granted.

3. Methods

This section describes the data, feature extraction, machine learning classifying models, and summary methods used for our results. We choose to use balanced datasets (*i.e.* those with approximately the same number of samples per class) for the sake of visualizations, though the results remain about the same with natural distributions. We used common strategies for feature extraction including term frequency inverse document frequency (TFIDF), n-gram, and PoS. Additionally, we compared different machine learning algorithms to see how they performed under a variety of conditions. We then summarized the text by using a reductive model.

3.1 Data & Preprocessing

We start by considering the Vox Media 2017 dataset [1]. Most authors have fewer than 50 articles, yet authors with more than 50 articles account for 91% of all articles published. Similar, most topics have fewer than 50 articles. There are also several articles written by multiple authors and some author's names are clearly pseudonyms, for example "A #NeverTrump Delegate". Also, many of the topics are related, which we deal with in Section 4.2. To deal with this skewed data, we construct two curated subsections of the data containing balanced number of articles per class (*i.e.,* author or topic). One contains 10 classes, each with 300 articles, the other contain 50 classes, each with 50 articles. Having the dataset balanced is useful for dissecting the results in the confusion matrices (Figures 3-4), but do not significantly improve overall accuracy. We also choose to ignore topics such as "Life", "Identities", and "The Latest" because we found that they tend to act as a miscellaneous category for Vox instead of focusing on a topic. We also filtered out the topics "Xpress" and "Vox Sentence" which tend to have very short articles, which makes them unsuitable for this task in addition to often being vague. This dataset has many favorable features such as being well curated for machine learning, including author and topic labels, and including inductive summary, which most other datasets such as *20 news organization* do not have.

3.2 Features

Machine learning models use features from the text to learn the class signatures. To this end, we extracted three types of features. First, we use word count and word bigram count. We also use word and word bigram TFIDF. We also use Natural Language Tool Kit's (NLTK) built in TreebankWordTokenizer and tagger to do PoS. We also limit the number of features in order to train the models more efficiently. We ignore features that are either very common or very rare as they are prone for bloating or overfitting, specifically by limiting features with term frequency between 0.01 and 0.99. We use the Random Forest (RF) model for feature importance evaluation. We also exclude all non-alphabetic characters besides spaces and periods. We exclude some features that are artifacts of the web embedding. Finally, we use the RF feature importance metric to look at what features are most important for distinguishing classes. Though we tried various features types, we found that n-gram term frequency performed best, while also not causing overfitting and use the same feature construction parameters for both authorship and topic.

3.3 Classifier Models

Discrete classification is a machine learning task with many classifiers readily available. We use several different learning algorithms and techniques as a comparative opportunity. We use naïve bayes (NB), decision trees (DT), RF, support vector classifiers (SV), and multi-layer perceptron neural networks (NN). These learning algorithms are supported by many open source libraries. We use a one vs. all (OvA) strategy with the NN to get slightly better results.

3.4 Text Summarization

To generate the text summaries, we used the NLTK sentence and word tokenizer and for removing stop words (*i.e.*, common words). Then, we tokenize at the sentence and word levels; filter out the stop words; and calculate the frequency of every non-stop word. We then score a sentence by the sum of the frequency score of the words, divided by the length of the sentence. We are then able to score the sentences and keep the sentences with high scores according to a threshold. Since this method filters for the most salient sentences and does not create new sentence structures or introduce new words or phrases it is an extractive rather than abstractive method. More complex summary methods could be applied to these texts, but we will consider that for future work. Informative summaries are generally around 20% the length of the original text[14], therefore we give about 10% margin on either side and remove 70%-90% of the original text. The threshold for doing this varies on the number of sentences and length of the document. To address this, we apply the summarization method several times until it converges to within the desired margin. If it cannot converge, we discard the document. The reason for not being able to converge is likely from sentences not falling into the threshold we set, which could be because the sentences are too long or the text is too short. The occurrence is rare and likely does not artificially inflate the results. The blurbs that are included in the dataset are on average (2.1+/-0.7%) the length of the article for our experiments, which is on the low end of acceptable for inductive summaries.

$$Sentence\ Importance = \frac{\Sigma(Word\ Frequency)}{Number\ of\ Words\ in\ Sentence}$$

Equation 1. Simple sentence importance calculation

3.5 Unsupervised Techniques

Text that is found in the wild is messy. The Vox Media dataset has the advantage of being well sorted and with pre-assigned labels, but even with this advantage, it has characteristics that make classification difficult. For example, many of the labels are closely associated and topics that vary in size and breadth. This is why initial research on this data focused on unsupervised approaches to topic clustering and lacked direct accuracy results as included in work. Related work showed that there are topics that emerge from the data such as, politics, entertainment, *etc* (see figure 2) and found that there are various numbers of clusters that have good coherence.[2] Other related work uses Latent Dirichlet Allocation (LDA) [17] to analyze how topics words compare to those of high importance to RF classification. LDA works by comparing word and topic distributions. By comparison, RF ranks the words by importance by finding the words which most separate topics.

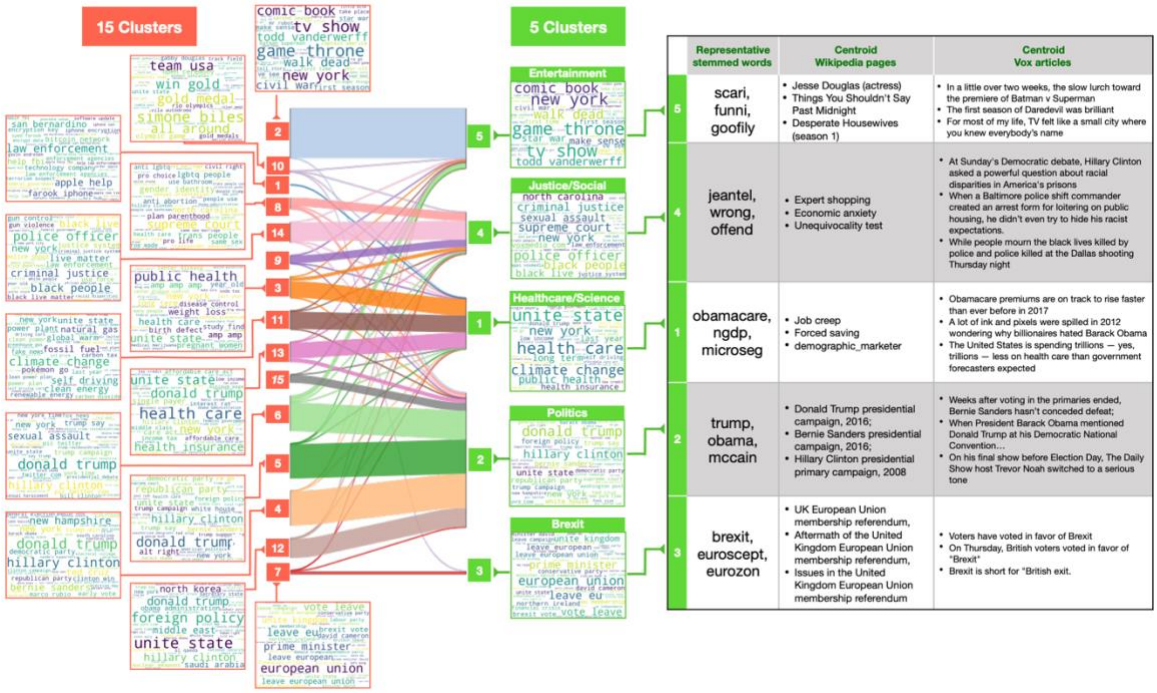

Figure 2. Sankey Diagram from [2] demonstrates an unsupervised approach clustering Vox Media articles by topic with 15 clusters to 5 clusters. Also included are root words and examples of wiki pages and Vox article titles.

4. Experiments & Results

The results of our work are a comparative analysis of author vs. topic classification with full text and summaries. For the author and topic comparison we use the same models, features, and dataset structure, though the individual articles may differ. Summaries were generated from the articles directly. We also include unsupervised learning to get an intuitive understanding of the data, such as scatter of article clusters and list of topic words. Finally, we consider how to handle problems with topic overlap and edge cases for authorship.

4.1 Authorship vs. Topic Classification

We start with stylometry. We can detect the style of an author statistically by doing the feature extraction as described previously. In the dense dataset, we trained with articles of the top 10 most prolific authors of Vox. With the dense dataset, we used 80% for training and saved 20% for testing, and attained up to 84% accuracy using a neural network with the OvA (NN_OvA); though the other methods also behave fairly well for this task. With the Sparse dataset we have some loss in signal but still strong considering there we are classifying 50 authors with 70% accuracy, whereas the baseline for guessing is 2%.

Table 1. Authorship attribution accuracy with various Machine Learning models on with n-gram word counts features.

Model	**Dense % Accuracy**	**Sparse % Accuracy**
Naïve Bayes	81	64
Decision Trees	53	30
Random Forest	74	51
Support Vector	74	32

Neural Network	83	51
Neural Network One-vs-All	**86**	**70**

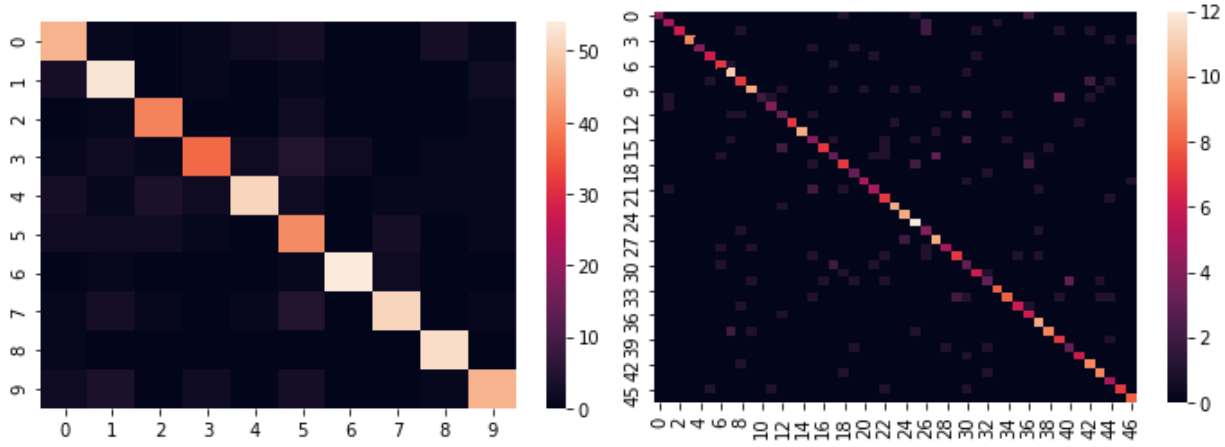

Figure 3. Authorship confusion matrix for dense using NN_OvA model (left) and sparse using NN_OvA model (right).

Topics can be classified with between 62%-74% accuracy. Therefore, given similar information, topics are 10% less accurate with dense information and 8% less accurate with sparse information. Looking at the confusion matrix of topics with dense information, it appears that one topic tends to dominate, and in the sparse dataset there appears to be two that were misclassified as each other. Whereas in the authorship case, the errors are more scattered. Additionally to compare this approach to other work with fewer number of topics, such as in [7] and fewer articles we were able to score 90% accuracy in distinguishing between "Politics & Policy", "Science & Health", "Culture", and "Business & Finance".

Table 2. Topic classification accuracy with various Machine Learning models on with n-gram word counts features.

Model	**Dense % Accuracy**	**Sparse % Accuracy**
Naïve Bayes	73	61
Decision Trees	55	45
Random Forest	70	62
Support Vector	65	38
Neural Network	72	53
Neural Network One-vs-All	**74**	**62**

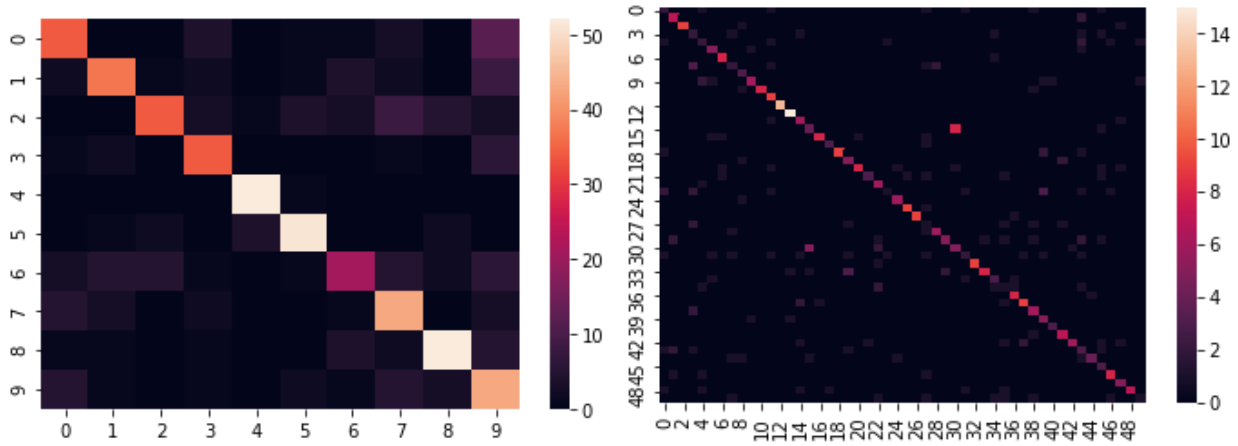

Figure 4. Topic confusion for dense using NN_OvA model (left) and sparse for dense using NN_OvA model (right).

We consider two kinds of summaries: inductive and informative. The Vox Media dataset comes with a short "blurb" for each article which makes for good comparison with our generated summaries. Each blurb is ~2% the length of the original text. The summaries we generate end up being ~17% the length of the original text. As can be seen in Table 3, summaries lose about 25% authorship signal and 20% topic signal. The scores that they get are still far above random guessing for both author and topic classification, though some useful information can be lost. This may be because the text becomes more general. So in terms of evaluating the quality of the summary, it clear that the longer summaries that we generate are better suited for learning models. This approach to summaries can be helpful if the amount of data is large by reducing it by 80% but the models work better with access to the full text. There should be continued exploration into abstractive summaries for this same task. The hope is that by abstracting information in novel ways (*i.e.*, not verbatim from the text) that salient information could be condensed more effectively.

Table 3. Classification Results on Summaries using NN_OvA model

Data Source	**Dense % Accuracy**	**Sparse % Accuracy**
Blurb for Authorship	19	8
Generated for Authorship	60	42
Blurb for Topic	21	5
Generated for Topic	53	41

4.2 Unsupervised Insights

The prior analysis focused on supervised learning with handcrafted labels as given by Vox Media organization. However, unsupervised learning can provide insights into trends within the data. We first consider how feature importance as learned from the RF compares to related words extracted by LDA. The LDA groups words by certain components and gives the top words for each component. The top 10 words for top 10 topics for each dataset and the feature importance from of the top 100 words from the RF. They share many of the same words with high importance such "Trump", "health", and "people". There are two potential issues we see from this, 1 (as addressed in Section 4.3), there are topics that overlap, 2 (as addressed in Section 4.4), authorship seems to be tied to topic in some way.

4.3 Adjusting for Topic Overlap

As we explored in the ways to address topic overlap, such as hierarchical and top-n approaches, the unsupervised language models may be more indicative of patterns within a of body of text. Therefore, we

provide the reader with some visuals of an unsupervised view of the data. We use principle component analysis to visualize the data based on author and topic and then use k-means clustering with 10 clusters to show how that fits the data.

One of the problems we noticed with the labeled topics is that some are general or closely related to other topics. To address this, we ease the classification by a top-n and a hierarchical approach for the sparse topic data. This is the case that makes the most sense because there are enough categories of things that could be conflated. Using the top-5 topics brings the accuracy from 62% up to 87%. To do the hierarchical model, we have to manually select topics of each group. Informed by [2], and descriptions online, we group 5 super-topics, each with a number of subtopics (see Table 4) and include 800-1000 samples each. Using the machine learning models we can get 84% accuracy.

Table 4. Five general topics and associated subtopics

General Topics	**Total # Articles**	**Subtopics**
Politics	5479	Politics and Policy, Politics, Mike Pence, Ted Cruz, Congress, Hillary Clinton, Marco Rubio, Donald Trump, Jeb Bush, Bernie Sanders, Mischiefs of Faction
Health	1173	Health Care, Infectious Disease, Obama Care, Science & Health
Environment, Technology & Business	1084	Energy and Environment, Grist, New Money, Apple, Transportation, Space, Business & Finance, Technology, Labor Market
Social Issues	848	LGBTQ, Identities, Race in America, Marriage Equality
Entertainment	1442	Books, Game of Thrones, Movies, Culture, Music, Episode of the Week, Star Wars, Reviews

This approach differs from the top-n approach because we had to manually choose groupings, whereas with top-n, each article may have a different top grouping and still score correctly. This approach has the advantage that one can specify the topic and subtopics but works at a comparable level for a much smaller range of topics and needs more data.

4.4 Stylometry via Intra-topic and Inter-topic Authorship Classification

We suspected that there may be confusion in the signal between topics and author. As mentioned in [18] there may be irrelevance by correlated features which, when under unfortunate circumstance, cause highly confident incorrect classifications. Their example uses rotating images in the MNIST data. The concern in our case is that the topic may be indicative of the author. After all, some authors specialize in topics so instead of style detection, maybe it is a sort of article detector. While our goal is authorship classification, it is not strictly style. But to address this, we also consider trying to detect style by containing samples from within a topic. We choose to run experiments for The Latest, Donald Trump, and Politics and Policy because they had enough authors with enough articles each to do comparable experiments. The topics had between 8-10 authors with 60-300 (see Table 5). For the experiment on "The Latest", which most resembles the dense dataset, it scores even higher, but this may be because it is a miscellaneous category. Whereas the experiment on "Donald Trump" yields 64%, maybe because it had fewer samples or because it was more specific. We also performed an intra-topic experiment where authors were allowed only one sample per topic. For this experiment we had 8 authors with 50 samples each and it scored 68% accuracy. It would appear that authorship is actually easier to detect within a topic, but it can be detected whether the author focuses on one topic or writes about many with consistent accuracies.

Table 5. This shows how well authorship stylometry works within topics and the #articles indicates the number of articles per author.

Topic	Number of Authors	Number of Articles per Author	Accuracy
The Latest	8	300	92%
Donald Trump	9	60	64%
Politics & Policy	10	60	81%

5. Limitation, Discussion, & Future Work

This work provides insight into common NLP techniques and tools for a large unexplored dataset. It shows what kind of accuracy to expect from a dataset in which the text was well edited but large and shows that state-of-the-art accuracy can be achieved for authorship and topics. We also suggest ways of dealing with topic overlap in new contexts and discuss handcrafted vs. naturally occurring groupings. We hope these results are helpful for other NLP researchers in the pursuit of linguistic knowledge and that future research use it to enlighten their search and find better ways to achieve similar goals. We also demonstrate that a reductive approach to text summarization retains both authorship and topic signal to some degree. However, other summarization approaches could be explored in this context for interesting results.

5.1 Discussion

We see that we can use topic and author signals to classify documents. One concern that was raised is how these signals conflate. It is my belief that they are inexorable intertwined with regard to authorship. An example of this concern is that if an author writes a lot about a specific topic, what is classifier picking up on? So, for example, in [19] they use stylometry to test for plagiarism using student academic papers as their corpora. But since academic papers are required to be novel and are usually about very specific topics, it is not clear that they are not picking up on authorship or topic similarity.

There has been some work on how to know whether or not to trust your classifier when there is a "data shift". Their method deals with classifying the MNIST dataset and rotating the images. However changing between perhaps non-independent classification, there may be no way to disentangle with certainty,[18] or one may need to be aware of out of distribution changes in the data.[20] However, with unaltered data, this is generally not a problem but is necessary for generative models adversarial models. That being said, it is unclear what is the degree of Vox's editing signature that is included in the signal. An interesting question is, if the topic signal would shift when the text was edited to imitate someone else's writing. Changing the phrasing of sentences can throw off these types of attribution. Tools like ParChoice [21] retain semantics while changing specific words. These types of adversarial should be considered for creating robust models. Another way of evaluating the semantics in these cases would be to make sure they still do well in the topic classification cases. If that fails it is likely that they are changing the meaning rather than the style.

5.2 Limitations

There are some ways that this work is limited. It focuses on only articles from Vox, but could be expanded to include articles from other news sources. It also only uses simple machine learning methods, but more advanced neural networks and architectures could be used. Additionally, we could use other methods for generating summaries. We also focus just on English texts, but could apply these techniques to other languages as well.

5.3 Future Work

We measure the efficacy of summary generation for machine classification contexts. The method we used, called extractive, is useful because it is fast, flexible and easy to use. Summaries can also be generated using different means [3], which might result in different or better outcomes depending on the task. Methods involving generation rather than reduction [20] showed that one can adjust for domain data and could be considered for generative methods. Perhaps this could be used to improve text generating GANs. Using these methods on sources with multiple label introduces an interesting multioutput problem.

In addition to improving tasks explored here, there are other interesting pursuits one could take with this data or similar data, to explore where the learning is transferable. There is the concern of various signals being present or biasing the text. Work related to privacy and anonymity are often a concern when it comes to identifying individuals. It is important to be aware that these methods are largely used as supporting forensic evidence rather than absolute truth. However, this could also be used for good if we can use it to debias text or use multiple texts to form a multisource summary.

6. Conclusions

This work explores the new and rich news article data set provided by Vox Media for the NLP community. We demonstrate that state-of-the-art classification approaches with off-the-shelf language and learning tools are well suited for news articles, even though they may have been edited. We provide direct comparison between style and topic features and show that author attribution can score between 70%-86% accuracy for groups between 10 and 50 authors and between 62%-74% for 10 to 50 topics. The topic accuracy gap can be compensated for, when considering topic overlap in grey areas such as comparing topics like political figures and general politics. We compare top-n and hierarchical topics and combing methods to increase the score to 87%. Additionally, we show that simple extractive summarization techniques retain both authorship and topic signal and show how this compares to human generated abstractive summaries.

Acknowledgements

We want to thank Vox Media for providing the data used in this work and for their commandment to adapting journalism to the digital age. Additionally, this work was reviewed and encouraged by the Army Research Laboratory and Funded by the Auerbach Berger Chair in Cybersecurity held by Spiros Mancoridis, at Drexel University

References

[1] Vox Media, Workshop for Data Science + Journalism (DS+J) (2017)

[2] Altuncu, M. Tarik, Sophia N. Yaliraki, and Mauricio Barahona. "Content-driven, unsupervised clustering of news articles through multiscale graph partitioning." arXiv preprint arXiv:1808.01175 (2018).

[3] Kiesel, Johannes, et al. "Semeval-2019 task 4: Hyperpartisan news detection." Proceedings of the 13th International Workshop on Semantic Evaluation. 2019.

[4] Bountouridis, Dimitrios, et al. "Explaining credibility in news articles using cross-referencing." SIGIR workshop on ExplainAble Recommendation and Search (EARS). 2018.

[5] Khan, Junaed Younus, et al. "A benchmark study on machine learning methods for fake news detection." arXiv preprint arXiv:1905.04749 (2019)

[6] Lee, Stephen M. "Variation in Political News." (2019).

[7] Suh, Yirey, et al. "A comparison of oversampling methods on imbalanced topic classification of Korean news articles." Journal of Cognitive Science 18.4 (2017): 391-437.

[8] Kaur, Kamaldeep, and Vishal Gupta. "A survey of topic tracking techniques." Int J 5 (2012).

[9] Hughes, James M., et al. "Empirical mode decomposition analysis for visual stylometry." IEEE transactions on pattern analysis and machine intelligence 34.11 (2012): 2147-2157.

[10] Mara, Michael. "Artist attribution via song lyrics." (2014).

[11] Caliskan-Islam, Aylin, et al. "De-anonymizing programmers via code stylometry." 24th {USENIX} Security Symposium ({USENIX} Security 15). 2015.

[12] Ramnial, Hoshiladevi, Shireen Panchoo, and Sameerchand Pudaruth. "Authorship attribution using stylometry and machine learning techniques." Intelligent Systems Technologies and Applications. Springer, Cham, 2016. 113-125.

[13] Abbasi, Ahmed, and Hsinchun Chen. "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace." ACM Transactions on Information Systems (TOIS) 26.2 (2008): 1-29.

[14] Tas, Oguzhan, and Farzad Kiyani. "A Survey Automatic Text Summarization." PressAcademia Procedia 5.1 (2007): 205-213

[15] Nenkova, Ani, and Kathleen McKeown. "A survey of text summarization techniques." Mining text data. Springer, Boston, MA, 2012. 43-76.

[16] Babar, S. A., and Pallavi D. Patil. "Improving performance of text summarization." Procedia Computer Science 46 (2015): 354-363.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research 3, Jan (2003), 993–1022

[18] Ovadia, Yaniv, et al. "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift." Advances in Neural Information Processing Systems. 2019.

[19] Ramnial, Hoshiladevi, Shireen Panchoo, and Sameerchand Pudaruth. "Authorship attribution using stylometry and machine learning techniques." Intelligent Systems Technologies and Applications. Springer, Cham, 2016. 113-125.

[20] Ren, Jie, et al. "Likelihood ratios for out-of-distribution detection." Advances in Neural Information Processing Systems. 2019.

[21] Gröndahl, Tommi, and N. Asokan. "Effective writing style imitation via combinatorial paraphrasing." arXiv preprint arXiv:1905.13464 (2019).