

ABSTRACT

Title of Document:

DEVELOPING QUANTITATIVE
METHODOLOGIES FOR THE DIGITAL
HUMANITIES: A CASE STUDY OF 20TH
CENTURY AMERICAN COMMENTARY ON
RUSSIAN LITERATURE

Robert Cai, Matthew Thomas Carr, Adam Elrafei,
Alexander Goniprow, Adrian Hamins-Puertolas,
Manpreet Khural, Andrew Li, Alexandra Winter,
Soumya Yanamandra, Dan Yang, Kay Zhang

Directed By:

Dr. Peter Lancelot Mallios, J.D.

Using scientific methods in the humanities is at the forefront of objective literary analysis. However, processing big data is particularly complex when the subject matter is qualitative rather than numerical. Large volumes of text require specialized tools to produce quantifiable data from ideas and sentiments. Our team researched the extent to which tools such as Weka and MALLET can test hypotheses about qualitative information. We examined the claim that literary commentary exists within political environments and used US periodical articles concerning Russian literature in the early twentieth century as a case study. These tools generated useful quantitative data that allowed us to run stepwise binary logistic regressions. These statistical tests allowed for time series experiments using sea change and emergency models of history, as well as classification experiments with regard to author characteristics, social issues, and sentiment expressed. Both types of experiments supported our claim with varying degrees, but more importantly served as a definitive demonstration that digitally enhanced quantitative forms of analysis can apply to qualitative data. Our findings set the foundation for further experiments in the emerging field of digital humanities.

DEVELOPING QUANTITATIVE METHODOLOGIES FOR THE DIGITAL HUMANITIES:
A CASE STUDY OF 20TH CENTURY AMERICAN COMMENTARY ON RUSSIAN
LITERATURE

By

Team POLITIC

(Political Opinion and Literature: Identifying Themes in International Commentary)

Robert Cai, Matthew Thomas Carr, Adam Elrafei, Alexander Goniprow, Adrian Hamins-
Puertolas, Manpreet Khural, Andrew Li, Alexandra Winter, Soumya Yanamandra, Dan Yang,
Kay Zhang

Thesis submitted in partial fulfillment of the requirements of the Gemstone Program
University of Maryland, College Park 2014

Advisory Committee:
Dr. Peter Lancelot Mallios, Chair
Dr. Maurine Beasley
Mr. Travis Brown
Dr. Piotr Kosicki
Ms. Nataliya Pratsovyta

Acknowledgements

All of us would like to express tremendous gratitude to our mentor, Dr. Peter Lancelot Mallios, for his incredible expertise and tireless motivation throughout all four years of our project. We are indebted to Mr. Travis Brown and Mr. Nicholas Slaughter, as well as the entire faculty at the Maryland Institute for Technology in the Humanities and all those involved with the Foreign Literatures in America Project, without whom this research would not have been possible. We thank the Gemstone staff and our librarian, Mr. Tim Hackman, for their enthusiastic support from day one, and extend a special acknowledgement to Dr. William Mallios and Dr. Ronna Mallios for their consultation in substantiating our results.

Table of Contents

Abstract.....	i
Cover Page.....	ii
Acknowledgements.....	iv
Table of Contents.....	v
List of Figures and Tables.....	vii
Chapter 1: Introduction.....	1
1.1 Team and Project Overview.....	1
1.2 Project Design and Focus.....	2
1.3 Significance and Limitations of Findings.....	3
Chapter 2: Literature Review.....	4
2.1 Political Uses of Canonical Literature.....	4
2.2 Readers' Guide Retrospective.....	4
2.3 20 th Century United States-Russian Relations.....	5
2.4 Data Mining.....	5
2.4.1 Datasets.....	8
2.4.2 Text Mining.....	8
2.5 Waikato Environment for Knowledge Analysis (WEKA).....	9
2.5.1 Preprocessing the Data.....	9
2.5.2 Classification.....	12
2.5.2.1 J48 Decision Tree Classifier.....	12
2.5.2.2 The Validity of Decision Trees.....	14
2.5.2.3 ROC Area.....	15
2.6 Topic Modeling.....	16
2.6.1 Latent Dirichlet Allocation.....	17
2.6.2 Topic Modeling Standards and Accuracy.....	18
Chapter 3: Project Design and Methodology.....	19
3.1 Scanning.....	19
3.1.1 Microfilm Scanning Protocol.....	22
3.1.2 Print Periodical Scanning Protocol.....	23
3.2 Optical Character Recognition.....	25
3.3 Annotations.....	27
3.3.1 Developing the Questions.....	28
3.3.2 Sample Size.....	28
3.3.3 Selection of Annotation Articles.....	29
3.3.4 Documentation of the Annotation Process.....	29
3.3.5 The Annotation Process.....	29
3.4 WEKA.....	29
3.4.1 Creating the ARFF file.....	30
3.4.2 Preprocessing and Filtering the Data.....	31
3.4.2.1 WEKA Customization Options.....	32
3.4.3 The Four Final String-to-Word-Vector Filter Configurations.....	34
3.4.4 J48 Decision Tree Classifier and the ZeroR Classifier.....	35
3.4.5 Tenfold Cross Validation.....	36

3.5 Topic Modeling.....	37
3.5.1 Creation of the Topics Spreadsheet.....	37
3.5.2 Adjustment of Topics.....	38
3.6 Regression Methodology.....	39
3.6.1 Factor Analysis.....	40
3.6.2 Stepwise Binary Logistic Regression.....	41
3.6.3 Topic Modeling Experiment 1.....	41
3.6.4 Topic Modeling Experiment 2.....	42
3.6.5 Topic Modeling Experiment 3.....	43
Chapter 4: Results and Discussion.....	44
4.1 WEKA.....	44
4.1.1 General Trends.....	44
4.1.2 The Most Successful Classifiers.....	47
4.1.3 Classifiers for Questions Difficult to Annotate.....	51
4.1.4 Classifiers with a Modest Improvement in Accuracy.....	54
4.1.5 Underperforming J48 Classifiers.....	55
4.1.6 Sentiment Analysis.....	58
4.1.7 Suggestions for Improvement.....	61
4.1.8 Concluding Remarks.....	63
4.2 Topic Modeling Experiments.....	63
4.2.1 Topic Modeling Experiment 1.....	63
4.2.2 Topic Modeling Experiment 2.....	69
4.2.2.1 Authors.....	69
4.2.2.2 Sentiment Analysis.....	73
4.2.2.3 Radical Politics as an Issue.....	74
4.2.2.4 Style of Author as an Issue.....	77
4.2.3 Topic Modeling Experiment 3.....	79
Chapter 5: Conclusions.....	90
5.1 Future Considerations.....	91
5.1.1 Annotations.....	91
5.1.2 Quantifying Foreign Policy.....	93
5.2 Final Remarks.....	95
Appendices.....	96
Appendix A: Timeline of United States-Russian Relations.....	96
Appendix B: Sample Alternative Spellings of Russian Names.....	105
Appendix C: Scanning and OCR Guidelines.....	106
Appendix D: Sample Annotation Question Evolution.....	125
Appendix E: Annotation Questions and Guidelines.....	126
Appendix F: Downloading WEKA and Generating an ARFF File.....	129
Appendix G: Preprocessing the Data and Using Machine Learning Algorithms.....	131
Appendix H: Partial Rotated Component Matrix for Factor Analysis Data (Factors as Columns).....	137
Appendix I: Results of All Four Decision Tree Configurations for Each Annotation Topic.....	138
Appendix J: Topic Modeling Experiment 1 Tables.....	141

Bibliography.....	143
-------------------	-----

List of Figures and Tables

Figure 1. Decision tree for labor negotiations dataset (Witten, Frank, and Hall 18)	7
Figure 2. String-to-word-vector menu with parameters of interest highlighted in red.....	31
Table 1. Default settings of unused parameters found in the string-to-word-vector filter.....	34
Table 2. The Four Configurations used by the String-to-word-vector filter to develop word dictionaries.....	35
Figure 3. Example of a topic generated from our database of articles.....	37
Figure 4. Selection of our topics with the number of topics set to ten.....	39
Figure 5. Selection of our topics with the number of topics set to 80.....	39
Table 3. The most and least successful configurations for each annotation question with their respective accuracy rates and ROC areas.....	45
Table 4. The most and least successful configurations for each sentiment analysis question with their respective accuracy rates and ROC areas.....	46
Figure 6. A segment of the J48 decision tree for the annotation question pertaining to radical politics.....	48
Figure 7. A segment of the J48 decision tree for the annotation question pertaining to the author’s national identification.....	49
Figure 8. A segment of the J48 decision tree for the annotation question pertaining to the style and artistry of the principal author.....	50
Figure 9. A segment of the J48 decision tree for the annotation question pertaining to socioeconomic class.....	51
Figure 10. A segment of the J48 decision tree for the annotation question pertaining to religion.....	51
Figure 11. A segment of the J48 decision tree for the annotation question pertaining to the similarities between Russia and the West.....	52
Figure 12. A segment of the J48 decision tree for the annotation question pertaining to the contrast between Russia and the West.....	53
Figure 13. A segment of the J48 decision tree for the annotation question pertaining to whether the principal author is a subject of debate.....	54
Figure 14. A segment of the J48 decision tree for the annotation question pertaining to race.....	55
Figure 15. A segment of the J48 decision tree for the annotation question pertaining to gender.....	55
Figure 16. A segment of the J48 decision tree for the annotation question pertaining to the mention of books or other literary works.....	56
Figure 17. A segment of the J48 decision tree for the annotation question pertaining to the mention of foreign place names.....	57
Figure 18. A segment of the J48 decision tree for the annotation question pertaining to the article writer’s gender.....	58
Figure 19. A segment of the J48 decision tree for the sentiment analysis annotation	

question for positive, negative, and neutral/mixed opinion articles.....	59
Figure 20. A segment of the J48 decision tree for the sentiment analysis annotation question for positive and negative articles.....	60
Figure 21. A segment of the J48 decision tree for the sentiment analysis annotation question distinguishing between positive and non-positive (i.e. negative and neutral/mixed opinion) articles.....	61
Table 5. Comparison of Model Accuracy for Multiple Years and Modifications.....	64
Table 6. Stepwise Model Variable Addition and Accuracy Increase.....	66
Table 7. SBLR Excluding 1920-1922.....	67
Figure 22. Predicted Article Values, Turgenev as Primary Author.....	70
Figure 23. Predicted Article Values, Tolstoy as Primary Author.....	71
Figure 24. Linear Model, Tolstoy as Primary Author.....	72
Figure 25. SBLR Model, Tolstoy as Primary Author.....	72
Figure 26. Linear Model, Sentiment Analysis.....	73
Figure 27. SBLR Model, Sentiment Analysis.....	74
Figure 28. Linear Model, Radical Politics as Issue.....	75
Figure 29. SBLR Model, Radical Politics as Issue.....	75
Table 8. SBLR Model Coefficients Output, Radical Politics.....	76
Figure 30. Linear Model, Style as Issue.....	77
Figure 31. SBLR Model, Style as Issue.....	78
Figure 32. Russian Years of Crisis.....	79
Table 9. Crisis Year Logit Models.....	81
Table 10. Predicted Outcomes for Experiment 3.....	82
Figure 33. Predicted Article Outcomes by Year.....	83
Figure 34. Mean Predicted Article Values by Year.....	83
Figure 35: Percent of Articles Estimated to be Written During Crisis.....	84
Table 11. Percent of Articles Estimated to be Written During Crisis, Grouped Years.....	85
Figure 36. Russia at War or in Domestic Conflict.....	86
Figure 37. Predicted Values, SBLR Crisis Model.....	87
Figure 38. Predicted Article Values by Year, Crisis Model.....	88

Chapter 1: Introduction

1.1 Team and Project Overview

Team POLITIC (Political Opinion and Literature: Identifying Themes in International Commentary) formed in the spring semester of 2011 through the Gemstone Program at the University of Maryland, College Park. Gemstone is a prestigious, interdisciplinary program that allows teams of undergraduate students to propose, design, and conduct four-year research projects. Dr. Peter Lancelot Mallios, Associate Professor of English and American Studies and Director of the Foreign Literatures in America Project (FLA), proposed a research topic regarding political science and foreign literature. Team member Alexandra Winter authored the formal proposal and the team grew to a total of eleven members. The project evolved over time to encompass the subjects of computer science, statistics, history, and journalism.

Our research addresses two major issues facing the humanities today. The first is big data, which consists of the enormously available amounts of information whose massive scale is beyond the compass of individual analysis alone. Such large amounts of information require innovative computer, quantitative, and technological tools to derive meaning from them. The second issue is globalization, or the active interdependence and permeation of world cultures over the course of the 20th century and into the 21st. Globalization is defined as processes attributed to the intercommunication of opinions, ideas, marketable goods, and culture globally on the international stage resulting in international integration. The second development places significance on comprehension and communication between world cultures, and also critically appreciates the ways that it may distort perception of other world cultures.

Our project uses new computer techniques of mining big data to generate an overview of a specific test case of American historical understanding of Russian culture. This method of

using quantitative tools to address questions and data generally associated with the humanities is an area of study within the emerging field of the digital humanities. Analyzing individual data points from big data may not produce significant results. The primary goal of Team POLITIC is to construct effective methods of processing big data as a whole, as in the case of tracking trends in globalization and public portrayals of other nations.

1.2 Project Design and Focus

In the interest of a realistic experimental design, the team focused the project's direction on a case study, specifically American perception of Russia through Russian literature during the initial era of modern America's and modern Russia's emergence as world powers, 1898-1938. The team decided to focus on this time period due to its contrasting political environments, the most notable example being before and after the 1917 Russian Revolution, which collapsed the Tsarist autocracy and created the Russian Soviet Federative Socialist Republic. Convenience with regard to copyright laws was also a factor, in that the time period allows for open access to data.

Research Question: To what extent can the digital humanities and popular data analysis tools analyze large datasets and generate useful data in qualitative fields of study?

Answering this question requires two steps. The first involves the compilation of large datasets, given that not all of the materials that might be analyzed through these means are readily available in a consistent, digital format. These digital humanities tools can only extract useful, clean data and trends from well put together, clean collections of data. The second step involves the analysis and statistical tests to reveal those trends and patterns.

1.3 Significance and Limitations of Findings

With the digital humanities still emerging as an area of research, contributions at this stage have the potential to shape the field and direct its advancement. Even inconclusive results and unsuccessful projects can prove to be important in showing what cannot be done to analyze qualitative big data. Our procedures and experiments are significant in that the methodologies created can be used as starting points for other researchers. Our results are also vital to a recent development in the humanities known as the globalization of American literary studies, given that “the mechanisms by which [differences between countries] are translated into literature have never been fully specified” (Corse, *Nations and Novels* 1279).

Both the narrow scope of our research material and time limited our project. Though the case study was a means through which the team was able to test the data analysis tools, the information revealed through our experiments only pertains to a section of Russian history. We were also unable to explore the wider foreign policy and socio-economic implications of our data due to limited time. Additionally, while the team was able to use and evaluate each of the tools involved in the project methodology, we only ran tests with each of the tools based on an expectation of how the tool would function. When tools failed to produce the types of data we expected, we did not experiment further to ascertain the best use of these tools in the digital humanities. Though these were our limitations, they provide further research opportunities and a basis for similar projects to be carried out.

Chapter 2: Literature Review

2.1 Political Uses of Canonical Literature

Political motivations shape a nation's literary canon, which subsequently projects that nation's identity. The idea of a national literature emerged in the late eighteenth century as a way of proving cultural independence on an international level (Corse, *Nationalism and Literature* 7). Studies suggest canonical or high-culture literature does not reveal how citizens perceive themselves, but rather how political elites in power want to envision their nation (ibid 74). These previous studies also turn to literary prizes and public recognition to define the most frequently appearing works as canonical or high-culture (Corse, *Nations and Novels* 1279). Unlike bestsellers or popular culture novels, canonical texts differ greatly between countries, as they are symbolic in value and not simply economic commodities. Theories of canon formation state that novels have to experience the conjunction of large sales and certain types of recognition to reach canonical status (Ohmann 206). This recognition refers to the critical reception of works found in publications that "carried special weight in forming cultural judgments," such as the *New York Times Book Review* and the *New Republic* (204). Scholars have not yet specified the ways in which upper classes or changes in political power have translated national differences into literature, or how nations have received and publicized the canonical works of other nations.

2.2 Readers' Guide Retrospective

The Readers' Guide Retrospective is a reference of articles published between 1890-1982. Its database "contain[s] comprehensive indexing of the most popular general-interest periodicals published in the United States and reflects the history of 20th century America" (Virginia Polytechnic Institute and State University). EBSCOhost, an online research service, recognizes the Reader's Guide as the "ultimate index of subjects in the popular press," as it

offers over three million articles from over 550 periodicals. Due to the nationwide circulation and popular readership of these publications, articles from the Readers' Guide Retrospective are more likely to reflect mass sentiment in the United States in the early 20th century than other smaller sources of news, which were not as influential at the time and tend to hold regional biases.

2.3 20th Century United States-Russian Relations

The team compiled a list of relevant political, economic, and religious events relating to United States-Russian relations throughout the late 19th and early 20th centuries. These events came from the following sources: *American-Russian Relations, 1781-1947* by William A. Williams; *The American Mission and the "Evil Empire: " The Crusade for a "Free Russia" Since 1881* by David Foglesong; *American-Russian Rivalry in the Far East* by Edward Zabriskie; *The American Image of Russia, 1775-1917* by Eugene Anshel; *The American Image of Russia, 1917-1977* by Benson Lee Grayson; *Russia, the Soviet Union, and the United States* by John Lewis Gaddis; *American Opinion and the Russian Alliance, 1939-1945* by Ralph Levering; *The Cambridge History of Russia* by Maureen Peerie; *Distorted Mirrors: Americans and Their Relations with Russia and China in the Twentieth Century* by Donald David and Eugene Trani; and *The Soviet Union: Internal and External Perspectives on Soviet Society* by Vladimir Shlapentokh et. al. (Appendix A: Timeline of United States-Russian Relations.)

2.4 Data Mining

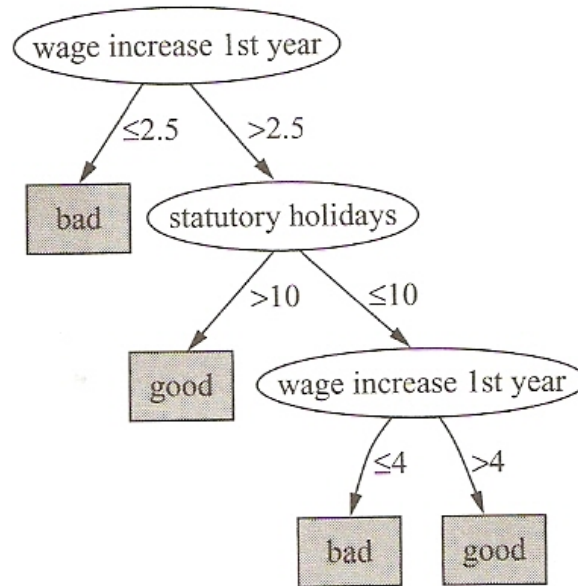
In today's world, access to data is overwhelming. Billions of books and articles are now available digitally through online research journals, popular magazines, and newspapers. However, making sense of such large volumes of digitized information is a complex task. Data mining offers methods of tackling large data. Data mining, or "the process of discovering

patterns in data,” uses computers with the necessary human input to analyze data found in datasets (Witten, Frank, and Hall 4-5). For our purposes, a dataset is a collection of data, and data is information that can take either a mostly numeric or nominal form. Thus, data mining allows one to use computers to extract useful, previously unknown information and patterns from a dataset that is simply too large to analyze by human endeavor alone (ibid 9).

The patterns revealed by data mining provide two important insights for that particular dataset. First, the patterns provide us with an understanding of the computer’s algorithms. Second, the computer can use the patterns to make predictions about new, but similar, data (Witten, Frank, and Hall 8-9). This predictive capability is one of the most significant aspects of data mining. The patterns are created by using machine learning algorithms. Since the generated patterns allow the computer to predict information about new data, the computer is considered to have learned new information. Hence, the term machine learning is used to describe the algorithms.

A classically cited example of data mining and machine learning is the Canadian labor negotiations dataset. This set includes many proposed labor contracts between management and labor and whether those contracts were accepted or rejected by both sides (Witten, Frank, and Hall 15). The dataset also includes the terms of each proposed contract such as the proposed wage increases for the first, second, and third year, the number of statutory holidays, and the proposed health plan benefits. For this dataset, a decision tree machine learning algorithm was used. This specific algorithm will be discussed later. But for now, the pattern produced by the decision tree can be seen in Figure 1 below.

Figure 1. Decision tree for labor negotiations dataset (Witten, Frank, and Hall 18)



In a decision tree, each circular node is considered a leaf, and the very top node is also called the root of the tree. The arrows emitting from each node signify the branching points of the tree. In this particular tree, “bad” signifies a contract rejected by management, labor, or both. “Yes” signifies a contract accepted by both management and labor. The decision tree algorithm allows users to better understand the computer’s thinking process. For example, if the wage increase for the first year is less than 2.5%, then the contract is rejected and is marked as “bad” by the algorithm. If the wage increase for the first year is more than 2.5%, then the algorithm considers another factor: the number of statutory holidays. More than ten days of holiday signifies an accepted contract, while fewer than ten days would require us to consider whether the wage increase in the first year is greater than, or less than and equal to 4%. Even though this is a relatively simple dataset, it acts as an adequate example to explain the previously mentioned concepts: through the use of a machine learning algorithm, the computer was able to data mine this dataset and extract patterns about the data without any human input in the process. Notice the pattern found that only the wage increase of the first year and the number of statutory

holidays were worth considering. The other information found in the dataset, such as health benefits, was deemed unimportant. Furthermore, we can easily interpret the generated pattern since it is in the form of a decision tree. Any computer with access to this decision tree and the appropriate software can now use it to predict whether both labor and management in Canadian companies would accept future contracts. However, it is extremely important to note that this decision tree may not always lead to the right prediction considering the fate of a contract.

2.4.1 Datasets

Data mining uses four terms to describe any dataset: instance, attribute, concept, and class. An instance is the individual examples in the dataset (Witten, Frank, and Hall 42). For example, in the labor negotiations dataset, each contract is an instance. Attributes are the information associated with the instance (ibid 49). Thus, the proposed wage increases of the first, second, and third year, the number of statutory holidays, and the proposed health plan benefits are the attributes for this dataset. The concept refers to the question we are trying to answer: Was the contract accepted or rejected? The class is the answer to that question (i.e. good or bad for the labor negotiations dataset) (ibid 39-40). In our Russian literary reception dataset, each document is an instance, the words associated with the documents are the attributes, the concept is one of the question we are trying to answer (i.e. is politics an issue?), and the class is the answer to that question (in our case, usually yes or no).

2.4.2 Text Mining

Our research falls under a subcategory of data mining called text mining. Text mining is an emerging field that involves “looking for patterns [specifically] in text” (Singh 315). Data mining differs from text mining in that the patterns in data mining are unknown to humans and machine learning algorithms are needed to extract these patterns from the dataset. However, in

text mining, patterns can potentially be found through human endeavor without machine learning tools. The major drawback is that such a process would take an enormous amount of time and effort (ibid).

For our research, we are focusing on a specific aspect of text mining called text classification in which documents are assigned to a predefined group by a machine learning algorithm (Singh 320). This type of text mining is considered a supervised process since it requires human knowledge to categorize a subset of texts from a larger text dataset into specific categories (ibid 322). The machine learning algorithm will then use the words found in those predefined text documents to develop a pattern, such as a decision tree, to categorize the remaining undefined texts found in the dataset. In our project, we use text classification to categorize each document in our entire dataset based on whether it fits under specific topics such as politics and religion. How do we approach text classification? To do this we make use of the machine learning workbench, WEKA (Waikato Environment for Knowledge Analysis).

2.5 Waikato Environment for Knowledge Analysis (WEKA)

WEKA, as described by its creators, “is a collection of state-of-the-art machine learning algorithms and data preprocessing tools” (Witten, Frank, and Hall 403). The development of WEKA was created at the University of Waikato in New Zealand in 1993 (Bouckaert et al. 2536). The current version with its graphical user interface was complete by 2005, and the software continues to be regularly updated (ibid 2537). For our project, we focus on two of WEKA’s features: its preprocessing tools and classification algorithms.

2.5.1 Preprocessing the Data

Preprocessing data refers to the tasks performed on a dataset before any machine learning takes place. As previously mentioned, in text classification processes, the machine learning

algorithm uses the words found in predefined documents (i.e. documents with their classes already assigned) in order to develop a pattern, such as a decision tree. The preprocessing tools found in WEKA allow us to deconstruct each document into the individual word components. This process is known as tokenization and is completed through the string-to-word-vector filter found in WEKA. Each word in a text file is also referred to as a word vector or a token (Witten, Frank, and Hall 329). Researchers justify deconstructing text documents into individual tokens through the bag of words approach in which text documents are viewed “as a sequence of words without considering [their] context...[or] words ordering” (Nuntiyagul et al. 32). In fact, using the bag of words approach is the most popular method in text classification studies (ibid).

When using WEKA’s preprocessing abilities to create a bag of words, many customization options are possible. These customizable options are the parameters found in the string-to-word-vector filter. For example, instead of building a bag of single words (unigrams), researchers have also used n-grams. N-grams are a consecutive sequence of words as they are found in the text. For example, each word in this sentence is a unigram. If we break down the sentence into groups of two words each, then each group would be considered a bigram. Hence, a group of 3 words is a trigram and so on. In some datasets, using n-grams have resulted in a significant increase in the accuracy of text classification by as much as 18% (Peng and Shuurmans 14). However, unlike Peng’s and Shuurmans’s work, other studies found the use of n-grams does not result in any significant increase in accuracy (Bekkerman and Allan 7). Thus, the possible benefit from the use of n-grams varies in different text datasets.

In addition to n-grams, we can also consider another parameter, the use of stemmers, when constructing our bag of words. When applying a stemmer to a text dataset, WEKA ignores prefixes and suffixes and only keeps the root of a word. Thus, the bag of words will consist of

the stems of words. For some text-based datasets, Sanderson and Watry found that stemmers cause a small increase in classification accuracy (78).

The use of n-grams and stemmers are not the only parameters found in WEKA's string-to-word-vector filter. WEKA also provides preprocessing options to remove words found on a stop word list from a text dataset. There is a default stop word list that includes common words that do not add any substantial meaning to a text such as "and," "the," "is," and "of." WEKA allows the use of a custom stop word list that removes words specifically tailored to the language found in a text dataset. Furthermore, the string-to-word-vector filter can be edited to include words in the bag of words that only appear a specific number of times in the text, or the words and their corresponding word frequencies. Just as the n-gram feature does not always result in an increase in accuracy as previously mentioned, there is no guarantee that any of these parameters will help increase our text classification accuracy. So how does a researcher decide how to preprocess his data into a suitable bag of words?

Witten, Frank, and Hall answer this question by stating that "text mining is a burgeoning technology that is still, because of its newness and intrinsic difficulty, in a fluid state... It is usually difficult to provide general and meaningful evaluations because the mining task is highly sensitive to the particular text under consideration" (389). In other words, they are saying no standard procedures exist for use in the text mining field. Due to the extreme variability that exists in different text datasets. For example, compare how language is used in a dataset of Tweets compared to a dataset of newspaper articles, there is no one suitable way to preprocess the data. Thus, it is left to each individual researcher to justify their use of preprocessing parameters and to experiment with many different configurations that they believe will work with their specific text dataset.

2.5.2 Classification

The bag of words created using WEKA's preprocessing tools is added in WEKA's word dictionary. The word dictionary is then used by machine learning algorithms known as classifiers in our text classification process to develop and extract patterns from the dataset that can be used to predict the class of future instances. More specifically, classification is when "the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples" (Witten, Frank, and Hall 40). WEKA contains over 100 classifiers, and thus researchers are left with a great deal of options to choose from (Bouckaert 2534).

2.5.2.1 J48 Decision Tree Classifier

Out of the hundreds of classifiers available on WEKA, there is one that is preferred by most computer scientists. A poll conducted by the International Data Mining Conference found that the most popular machine learning algorithm is the C4.5 classifier (Witten, Frank, and Hall 375-376). This classifier is found as the J48 decision tree classifier in the WEKA workbench. Decision trees, as mentioned earlier, can easily be interpreted by researchers and thus allow one to understand the patterns found in data. Due to these two factors (its popularity among experts and its interpretability), we chose to use the J48 classifier as our machine learning algorithm for this project.

When constructing a decision tree, the J48 classifier uses the divide-and-conquer approach: the machine learning algorithm considers which attribute to use as the root of the tree and which attributes to branch on as it moves down the tree (ibid 99). To understand how the classifier selects attributes to use, we must consider two concepts, the impurity of a subset of data and its entropy. A pure subset of data is one in which "all the instances belong to the same class" (Croce 26). For example, consider the two classes, good and bad, from the labor

negotiations decision tree. A completely pure subset of data is one in which all instances are either marked as good or bad. An impure subset is one in which there is a mixture of good and bad classes. Entropy is the measure of impurity found in the complete dataset or a subset of the data (ibid 27). The original entropy of an entire dataset, $H[D]$ can be defined mathematically as:

$$H[D] = - \sum_{j=1}^C P(c_j) \log_2 P(c_j)$$

where D is the dataset, C is the class of interest, and $P(c_j)$ is the fraction of instances in that class. Notice that the purest dataset of solely instances from one class will have an entropy value of 0 (ibid 28). For example, consider a hypothetical dataset in which 75% of the instances are of one class, and the remaining 25% are of another. Then the entropy value equals:

$$H[D] = -0.75 \log_2(0.75) - 0.25 \log_2(0.25) = 0.6226$$

Now, if we consider the presence or absence of a specific attribute (in text mining this is a word vector), the original dataset will be subdivided into subsets of data (i.e. the tree branches). The sum of entropy of these subsets of data, $H_{A_i}[D]$, can be defined mathematically as:

$$H_{A_i}[D] = \sum_{j=1}^v \frac{D_j}{D} H[D_j]$$

where the dataset, D , is divided into v subsets after branching from the attribute, A (ibid 29). The J48 classifier calculates the $H_{A_i}[D]$ value for every attribute in the text. The entropy of the original dataset and the entropy when considering an attribute are related together through a concept known as information gain which is defined as:

$$gain(D, A_i) = H[D] - H_{A_i}[D]$$

where the gain is the difference between the entropy of the dataset and the entropy of the dataset after branching at a specific attribute. Recall that the $H_{A_i}[D]$ approaches 0 as the purity of

the subset increases. Thus, after selecting an attribute, the greater the information gain value, the greater the purity of the subsets after the tree branches. Recall that the purpose of the decision tree is to classify the instances into specific classes. Therefore, a large information gain value means that the attribute used to branch the tree was successfully able to classify many instances. Thus, WEKA calculates all information gain values for every word in the word dictionary and selects the one with the highest value as the root of the tree (Croce 31). This process repeats itself in selecting an attribute for each node until all instances are classified (ibid 45). The accessibility of WEKA lies in that all these mathematical concepts are already built into the J48 classifier, and thus a researcher only needs a general understanding of the divide-and-conquer approach to understand the outputted decision tree.

2.5.2.2 The Validity of Decision Trees

The validity of decision trees created by WEKA's J48 classifier is primarily assessed by its accuracy rate (Witten, Frank, and Hall 150). The accuracy refers to the number of correctly classified instances divided by the total number of instances. In other words, the accuracy rate reveals the percentage of instances the classifier correctly categorized. In fact, choosing the classifier that produces the highest accuracy rate is "quite sufficient in many practical applications" (ibid 156). Also, published research articles tend to compare the accuracy rate of different classifiers, or the accuracy rate of the same classifier using different word dictionaries, with the highest accuracy rate being declared the best classifier. For example, in her study in classifying email as spam or non-spam messages, Lakshmi and Radha use the accuracy rate to distinguish between different classifiers (2786).

Another important consideration in assessing the validity of decision trees is determining what is considered a good accuracy rate. This question is primarily answered by comparing the

classifier's accuracy rate to the baseline's accuracy rate. The baseline accuracy rate is acquired when a classifier predicts the majority class for each instance (Witten, Frank, and Hall 377). For example, consider a hypothetical dataset with 100 instances, 70 of which are of class A, and 30 are of class B. The baseline accuracy rate is 70% because if we predicted the majority class (in this case, class A) for each instance, we would calculate a rate of 70/100. Thus, a classifier that produces an accuracy rate higher than baseline has predictive qualities since it performs better than when one simply predicts the majority class.

However, there is no specific percentage to qualify a good accuracy rate. For example, Lakshmi and Radha were satisfied with their accuracy rates of over 90% in classifying emails as spam or non-spam messages (2786). On the other hand, Lee et al. were content with an accuracy rate of 70.96% in their study of classifying Tweets into 18 different categories. They justified their stance by referencing the difficulty of working with Tweets due to the use of abbreviations and the presence of 18 possible categories. They were also impressed in that their accuracy rate was 3.68 times greater than the base line's rate (256). Another example is the research of Anta et al. In this study, the researchers only managed to achieve an accuracy rate of 58.45%. Yet, they still justified the publication of their results citing that they were among the first to attempt to use classifiers generally used on English language text on Spanish language Tweets (Anta et al. 51-52). Naturally, all researchers aim for the highest accuracy rate possible, but in reality, their contributions to the field stem from how they preprocess their text, interpret their results, and their suggestions for improvement.

2.5.2.3 ROC Area

By definition, the baseline rate has an area under the Receiver Operating Characteristic (ROC) curve of approximately 0.5. The ROC curve is a graphical representation that relates the

true positive and false positive rate of a classifier (Bradley 1145-1146). A true positive is one in which a classifier predicts the true class of an instance. A false positive occurs when a classifier predicts the wrong class of an instance. More specifically, the area under the ROC curve, the ROC area, signifies “the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance” (Majnik and Bosnik 6). In other words, an ROC value that equals 1 is one in which all classes are true positives (i.e. the classifier accurately predicted the class of each instance), while an ROC value that equals 0 is one in which only false positives were observed. In addition to the baseline accuracy rate, a classifier that predicts the majority class will have an ROC area of approximately 0.5 (ibid 7). Thus, an ROC area that is greater than 0.5 represents a classifier that has predictive qualities and is statistically significant with the best classifiers approaching an ROC area of 1.

The classification accuracy rate and the ROC area are two factors that will help any researcher find the best classifier for their particular dataset. Witten, Frank, and Hall state, “Experience shows that no single machine learning scheme is appropriate to all data mining problems. The universal learner is an idealistic fantasy” (403) Thus, previous studies only provide guidance to choosing the parameters of a classifier, and should not be completely imitated. Text mining is an “experimental science” that will involve a great deal of testing by trial and error to achieve adequate results (ibid).

2.6 Topic Modeling

In his introduction to topic modeling in the *Journal of Digital Humanities*, David M. Blei provides the definition as “a suite of algorithms to discover hidden thematic structure in large collections of texts, [the results of which] can be used to summarize, visualize, explore, and theorize about a corpus” (n. pag.). The term “topic” denotes a group of related terms that tend to

occur together, represented by distributions over the vocabulary used throughout the corpus. The words with the highest frequencies appear at the beginning of the generated lists, and are the most relevant to our interpretation of what the topic conveys. Topic modeling requires a large corpus, a general understanding of the data, a tool to carry out the topic modeling, and some way of analyzing the findings (Brett n. pag.)

2.6.1 Latent Dirichlet Allocation

The simplest form of modeling a topic is latent Dirichlet allocation (LDA). LDA assumes each corpus contains a fixed number of topics and each article or file will contain each of those topics to some extent (ibid). The algorithm generates the topics and distributions independently without assigning meaning to words or topics. Lisa M. Rhody of the Maryland Institute for Technology in the Humanities describes topic modeling, specifically LDA, as “generative, unsupervised methods of discovering latent patterns in large collections of natural language text: generative because topic models produce new data that describe the corpora without altering it; unsupervised because the algorithm uses a form of probability rather than metadata to create the model; and latent patterns because the tests are not looking for top-down structural features but instead use word-by-word calculations to discover trends in language” (n. pag.).

LDA models temporal relationships among topics through the use of an extension (Hall, Jurafsky, and Manning). Blei and Lafferty’s Dynamic Topic Model represents topic distributions along a normal curves generated using the preceding year’s distribution, and allows for control of what constitutes the topic from year to year (14). Wang and McCallum’s Topics over Time Model keeps the topics constant, but word co-occurrence relationships are dynamic and time is continuous, rather than discretely measured in increments (425).

2.6.2 Topic Modeling Standards and Accuracy

While no standard for choosing the number of topics for the tool to generate exists, cross-validation is one option (Taddy, 1186). Cross-validation involves repeatedly generating different numbers of topics to the data and interpreting which number seems to capture the themes most accurately. Statistically proving the significance of results generated with this method is difficult, and due to the trial-and-error nature, it is not easily scalable. However, cross-validation is the most popular choice within the field of digital humanities at this time (ibid).

Due to the endogenous nature of topic modeling, the tool cannot create error unless the error was in the dataset construction itself. The topic models are not manipulated interpretations, but rather internal reflections of all the data. Error can occur within human interpretation of the topics or in choosing the number of topics. Blei and Lafferty state that they “conclude with a word of caution. The topics and topical decomposition found with LDA and other topic models are not ‘definitive.’ Fitting a topic model to a collection will yield patterns within the corpus whether or not they are ‘naturally’ there. Rather, topic models... provide a summary of the corpus that is impossible to obtain by hand... [and] may yield connections between and within documents that are not obvious to the naked eye, and find co-occurrences of terms that one would not expect a priori” (17).

Chapter 3: Project Design and Methodology

Our compiled database consists of all articles within the Readers' Guide Retrospective that pertain to Russian literary figures and Russian works of literature. Alternative author name spellings were researched to account for variations in English interpretation, and a sample of these differences can be found in Appendix B: Sample Alternative Spellings of Russian Names.

3.1 Scanning

All articles of interest in our Reader's Guide Retrospective database originated from hardcopy sources. We scanned these sources into a digital format for compatibility with our software for data analysis. The scanning process was defined and developed by Nicholas Slaughter, a member of the Foreign Literatures in America Project here at the University of Maryland, College Park. The Scanning sub-team, Nicholas, and other FLA associates used the scanning protocol (Please see Appendix C: Scanning and OCR Guidelines.) to digitize the entire database of online articles. Although different members of the team scanned different articles, we strictly adhered to the procedures and parameters in the scanning protocol, thus ensuring that the resulting articles would be uniform in image quality and scale.

Materials must be prepared before any scanning can begin. First, a USB flash drive or other storage device is necessary for backing up the scanned files. Second, a soft microfiber glass or lens cloth assists with cleaning the scanning surfaces. Third, a cleaning solution for the glass surfaces of the microfilm and book scanners is required. A writing utensil, preferably a pen, is useful for making notes and marking the Assignment Sheet. Lastly, although bringing a ruler is optional, it is highly recommended to enhance the accuracy and quality of scans, especially for print sources.

We strictly adhere to standardized file formats and file naming systems to facilitate uniformity and clarity between scanners and other team members. Scanned articles were saved as .TIF files. This file format is the largest data format for pictures, ensuring that our scans were saved in high quality. The .TIF files were saved using LZW or ZIP compression for “loss-less” compression as well in order to reduce the file size. Articles were saved in grayscale at 600 DPI (Dots per Inch). File naming followed this basic format:

Fla-“Research Category”-“User”-“Document #”-“Page #”“A/B”.TIF

Example Filename: Fla-Tolstoy-ayl1-0001-001A.TIF

“Fla” is the heading of each file and stands for the Foreign Literatures in America Project.

“Research Category” is the topic of interest, and is typically the last name of the Russian author

of whom the article discusses. “User” is the initials of first, middle, and last name of the

individual who scanned the article followed by the number one. “Document #” labels this

particular article page as belonging to a series under the specific Research Category. The first

article is labeled “0001,” the second as “0002,” and so forth. “Page #” labels the pages of each

article and always begins with “001” for the first page. “A/B” after the page number further

distinguishes the file as a cropped or uncropped image. “A” signifies a file cropped for the article

of interest, while “B” signifies an uncropped page in its original form. Lastly, the file extension

is always .TIF to follow the file format procedures.

Records for scans are kept in two formats: an online spreadsheet and Assignment Sheets.

The spreadsheet is named Scanning #-Name.xlsx, and uploaded to the SugarSync Assignment

Sheets folder. The following information must be entered for each scanned file:

- File Name
- Page #(s)
- Pages in Document
- Main Title
- Sub Title

- Alt Title
- Descriptive Title
- Author
- Placement in Publication
- Publication
- Volume
- Issue/Number
- Date (Month/Day/Season)
- Year
- Publisher
- Publisher Location
- Date Acquired

All fields are identical for all files in a single document (article) with the exception of “File Name” and “Page #(s).” Records must be made and maintained after each scan in order to document progress and allow for any team member to quickly obtain the details of origin of each file in the database.

The basic scanning protocol follows a series of defined steps. Prior to scanning, the Assignment Sheet is downloaded from our SugarSync database and printed. The Assignment Sheets were one page documents created by Mr. Nicholas Slaughter that defined the location, source type, source name, library call numbers, and page number(s) of the articles of interest. Each Assignment Sheet typically contains around several journals and a section for notes on scan quality and results. Completed Assignment Sheets are uploaded to the Completed Assignments folder on SugarSync. A new folder is created with the assignment title and the scanner’s last name. All scanned files are copied along with the spreadsheet into this folder. The location of the physical documents is listed on the Assignment Sheet. The locations we obtained these documents from are: McKeldin Library and Hornbake Library at the University of Maryland, College Park, and the University of Maryland, Baltimore County. An order was placed for off-site sources through McKeldin Library. Transfers took approximately one to two business days,

and were temporarily kept behind the Circulation Desk at McKeldin Library. Periodicals were returned to the Circulation Desk after scanning.

There are two types of sources that we scanned to obtain the articles of interest: Microfilm Reels and Print Periodicals. McKeldin Library's electronic catalog can be utilized to locate these sources. Microfilm sources are articles contained on reels of film located in the back past the Circulation Desk of McKeldin Library. Each reel can be found in a labeled box. Some boxes were unlabeled. If identified, the correct call number was written in pencil on the box. Microfilm must be scanned using special microfilm scanners located right next to the reel boxes. Library policy limits the number of reels that can be used to five at one time. Print Periodicals are articles contained in bound books located in the Periodical Sections (2F) of McKeldin and Hornbake Library as well as on off-site locations. It is helpful to borrow a cart to transport the books, as they are large and bulky. Many periodicals are aged and fragile, so great care must be taken when handling them. Periodical sources can be scanned at several of the large scanners in McKeldin Library. The main scanner of use is situated on the first floor to the immediate left of the Front Desk upon entry. There are additional scanners on the second floor of the library. An overhead scanner may be used in place of a conventional one if a book is too damaged and fragile.

3.1.1 Microfilm Scanning Protocol

After obtaining the microfilm reel boxes of interest, the scanner logged into the Scanning Station computer with his UMD Directory ID and password. The USB flash drive or other storage device is inserted into the computer. The scanner glass surface is sprayed with cleaning solution and wiped down with the glass or lens cloth to reduce dust or any other debris that could interfere with image quality. This cleaning step was repeated throughout the scanning process to

eliminate any new dust or debris that settled on the scanner surface. Upon successful login, the PowerScan 2000 program was opened from the desktop. The Medium: 35mm Microfilm button was clicked. The scanner then proceeded to set the Scanning Resolution to 600 DPI from File to Set on the drop-down menu.

The mechanical scanning portion of the microfilm protocol is as follows. The microfilm reel must first be physically loaded onto the scanner. The glass tray is pulled out and the microfilm placed on the spool. The end of the microfilm is weaved through the wheels, past the scanning surface, and attached to the second spool. For navigation, the button on the bottom of the screen in PowerScan 2000 can be used to rotate the microfilm reel or to correct the page and orientation. Super-fast-forward and Super-fast-rewind can be clicked when the glass tray is all the way out. Regular-fast-forward and Regular-rewind can be clicked when the glass tray is all the way in. These four options are normal forward and rewind commands, but differ in their speed with Super being faster than Regular. Lastly, there is also an option for Page-by-Page scrolling that can be accessed from the screen.

The digital scanning portion of the microfilm protocol is as follows. The Auto Adjust button can be clicked to fix lighting and focus on the image. After adjusting the green cropping box on the screen, the page can be scanned by clicking the Scan button. Only relevant areas were scanned. The cropping option allowed scanners to minimize blank space and remove unnecessary information. Two scans are made: a cropped scan containing only the article of interest (labeled "A") and an uncropped scan of the full page (labeled "B"). To save the scanned file, the options: File, Scan to Drive #1, and Save As must be selected on the drop-down menu of the program. Scans were saved on the computer first, then transferred to the flash drive or another source for efficiency. Saving scans directly onto the flash drive is slower than the

previous step. As a final record, all bibliographic information was recorded into the online spreadsheet. The Assignment Sheet was updated with relevant notes on scanning progress and quality, in addition to any title, page, date, or other discrepancies.

3.1.2 Print Periodical Scanning Protocol

After obtaining the necessary print periodicals, the scanner logged into the Scanning Station computer with his UMD Directory ID and password. The USB flash drive or other storage device is inserted into the computer. The scanner glass surface is sprayed with cleaning solution and wiped down with the glass or lens cloth to reduce dust or any other debris that could interfere with image quality. This cleaning step was repeated throughout the scanning process to eliminate any new dust or debris that settled on the scanner surface. The EPSON Scan program can be opened from the desktop. The initial settings should be grayscale, and 600 DPI resolution. Make sure the text is not enhanced.

The mechanical scanning portion of the print periodical protocol is as follows. The periodical was opened to the page containing the article of interest as outlined by the Assignment Sheet. Pages to be scanned were placed on the cleaned scanning surface. Scanning can be performed with either the cover closed or open. The orientation and scanning dimensions must be defined in the EPSON Scan program. It is recommended to scan two pages as one file to be efficient. A ruler is useful to make more accurate measurements. A Custom Size of 12” Height and 11.7” Width is recommended for scans, but not required. Only important, relevant areas are scanned. Blank space is minimized.

Once the periodical is oriented correctly and held down firmly on the scanning surface, the Scan button is clicked. Two scans are made: a cropped scan containing only the article of interest (labeled “A”) and an uncropped scan of the full page (labeled “B”). To save the file, File,

Scan to Drive #1, and Save As are selected from the drop-down menu. The resulting scan file should be saved in .TIF format with LZW compression to save memory space. Scans should be saved to the computer first, then transferred to a flash drive or uploaded to email to save time. As a final record, all bibliographic information was recorded into the online spreadsheet. The Assignment Sheet was updated with relevant notes on scanning progress and quality, in addition to any title, page, date, or other discrepancies.

3.2 Optical Character Recognition

Optical Character Recognition (OCR) is a process by which a computer program converts text from a scanned image into readable characters. It does not perfectly detect every character of the text, but accuracy improves when supervised with user input. The computer used for OCR is located in the Foreign Literatures in America Office on the first floor of Tawes Hall, Room 1202.

Once logged into the computer, ABBYY FineReader 11 is opened from the desktop. The Master Spreadsheet, which contains all of the files that have been and have yet to be run through OCR, is located through our file sharing site, SugarSync, under Shared Folders, Foreign Literatures in America, Russian Authors Initiative, and Assignment Sheets. The Spreadsheet is titled "New Files to OCR.xlsx." As there were multiple members of the team conducting OCR throughout the duration of the project, the Master Spreadsheet must be downloaded before the OCR session, updated with the new files that have been run through OCR, and uploaded over the old document on SugarSync at the end of the session.

Open Image/PDF is clicked, and the file to OCR is selected from the OCR Workspace folder on the desktop. It is important to note that if the page is a picture only, it must be labeled as such ("Picture Only") in the Master Spreadsheet. Only regular files and "A" files are run

through OCR. Picture Only and “B” files are skipped, but noted in the Master Spreadsheet. A selection area must be made prior to OCR. The text should automatically be encased in green selection boxes by default. In most cases, this is fine for reading, however alterations must be made depending on the area selected. The Text button selects the text of interest by highlighting boxes around it. Individual columns in separate selection boxes are read from left to right, top to bottom. The program reads the text in the order that the selection boxes are made as well. To reorder the selection boxes, one can click the Area properties tab and change the value in the Area # box. All texts including footnotes and captions must be included in the reading. However, the main body of the text was selected so that it was read first. The Add Area and Cut Area buttons are used to alter rectangle selection areas in case the text is not perfectly rectangular. Selection boxes can be formed in different polygonal shapes. The Delete key is used to remove extraneous marks, pictures, and selection areas.

To run the text through OCR, the Read button must be clicked. After reading, the user must visually check that the text is read properly, and if not, make the changes accordingly to the selection areas and Read again. The results are saved in three different formats: “.TXT” (OCR-A-), “.DJVU” (OCR-B-), and “FineReader Document” (OCR-C-) as previously stated. To increase efficiency, it is simpler to save the .TXT file and .DJVU file in the same folder: OCR A&B, and the FineReader Document in a separate folder: OCR C. Both folders are located on the desktop. The files are then moved to their respective permanent folder destinations in the OCR Workspace at the end of each session. The user must ensure the same number of OCR A, B, and C files are in their respective folders, and resolve the issue if a discrepancy arises.

OCR files are saved in three formats to maximize reading and conversion accuracy. “.TXT” is a plain text format. “.DJVU” is an editable copy. It produces an editable, slightly less

precise facsimile of the text on the original image. “FineReader Document” is an exact copy. It produces an un-editable, precise as possible facsimile of the text on the original image. Files are named in the following format:

“File Type”-Fla-“Research Category”-“User”-“Document #”-“Page #”.“File Format”
“File Type” is based on the file format. “OCR-A-“ is the prefix for .TXT files, “OCR-B-“ for .DJVU files, and “OCR-C-“ for FineReader Documents. “Fla” is the heading of each file and stands for the Foreign Literatures in America Project. “Research Category” is the topic of interest, and is typically the last name of the Russian author of principal discussion in the article. “User” contains the first, middle, and last initials of the team member who converted the files followed by the number one. “Document #” labels this particular article page as belonging to a series under the specific Research Category. The first article is labeled “0001,” the second as “0002,” and so forth. “Page #” labels the pages of each article and always begins with “001” for the first page. The File Format is the extension based on what type of file it is: .TXT, .DJVU, or FineReader Document.

Throughout the OCR process, our Master Spreadsheet, “New Files to OCR.xlsx,” required constant updates with the following information: file name read, date of reading, username, and notes. The file name is preset and is already written in the Master Spreadsheet. The username is the same acronym as the one used for scanning and all other SugarSync uploads, unique to each individual team member. The notes are optional, and describe the OCR results and details about image quality and content, such as “Just a picture.” or “Text unreadable.” The files are uploaded into the Completed OCR folder inside the Completed Assignments folder on SugarSync. The Master Spreadsheet is uploaded into the Assignment

Sheets folder on SugarSync. (Please see Appendix C: Scanning and OCR Guidelines for a complete guide to the procedures.)

3.3 Annotations

One of the goals of our research was to test machine learning. Specifically, we wanted to explore whether machine learning could automate annotating large collections of documents, and whether a computer could objectively analyze author sentiment. Ultimately, our goals were to investigate the trends in themes and sentiment, and relate them to historical context. To accomplish this, we first had to develop a set of questions to answer about our database of articles and then to individually annotate a sample of documents to provide the computer with data from which to learn.

3.3.1 Developing the Questions

The annotation sub-team met with our mentor weekly to discuss the phrasing of these questions. Our main objectives for the questions were to provide insight to the content in the article and the article author's opinions on the Russian literary figure or work of literature.

Our questions underwent a series of revisions. We initially saw a great deal of discrepancies in how annotators would answer the questions. As a result, we moved toward making our questions more explicit to best eliminate human biases in interpretation. For example, a question with potentially variable answers such as, "Is religion an issue/topic in this article?" became "Is religion ever explicitly referenced in this article?" If any key words such as a mention of an organized religion or a ritual associated with an organized religion were found within the article, then a "yes" would qualify for this question. (Please see Appendix D: Sample Annotation Question Evolution.) After a number of revisions, we developed clear identifiers for

what would constitute “yes” or “no” answers with regard to mentioning a variety of social topics. (Please see Appendix E: Annotation Questions and Guidelines.)

3.3.2 Sample Size

Once we finalized our annotation questionnaire, we created the dataset of annotated articles. As suggested by our advisor from MITH, Mr. Travis Brown, a compilation of 150 annotated articles would be a generally acceptable amount of data to serve our purposes of effectively utilizing machine learning technology. We were able to compile 241 annotated articles.

3.3.3 Selection of Annotation Articles

The articles picked for annotations were selected based on availability. The generation of our database of articles via scanning and OCR was occurring at the same time as our annotation process. Thus, the annotation team could only choose articles that had already been processed by the scanning and OCR team.

3.3.4 Documentation of the Annotation Process

We completed the annotation process online using Google Documents and Forms. While reading an article, the annotator would answer the questionnaire and submit it via Google Documents, which would then automatically compile the answers into an Excel spreadsheet that allowed us to easily view completed annotations.

3.3.5 The Annotation Process

Members on the annotation team formed pairs, each having an assigned set of articles. Each member would read the assigned articles and answer the questionnaires individually. When both members of the pair completed their separate annotations, they would compare their responses and submit a finalized set of agreed upon answers. This was done to reduce human

errors such as missing a reference to religion due to lack of familiarity or simple misunderstandings a word. If the members were unable to come to an agreement for any particular question, the article in question would be deferred to our mentor, who held the tiebreaking vote.

3.4 WEKA

The WEKA workbench allows us to produce a J48 classifier for each annotation question based on our 241 annotated articles. Based on this classifier, we can classify all the remaining non-annotated instances in our Russian literary reception dataset. This process would be inefficient if WEKA was not used. Every single article in our dataset of over 1100 documents would have to be accurately and individually annotated by at least two annotators. WEKA simplifies this process by requiring only an annotated subset of the larger dataset. Then, by using this information, WEKA creates a classifier that is used to predict the classes for the remaining non-annotated articles.

If the classifier produced by WEKA is a successful one with a high accuracy rate, then the possibilities are enormous. With all the articles annotated, we can analyze the distribution of articles for each class over time. For example, we can determine the percentage of articles that discuss radical politics and how the percentages changes over time. We can also compare the distributions of articles for different classes over time and determine whether historical events, such as the Russian Revolution, relate to which classes are most prevalent during a specific year or time period. Furthermore, we can analyze the decision trees to understand the thinking process for each classifier. Thus, WEKA may allow us to reveal patterns about Russian literary reception that may have been missed if computer tools were not used. In order to develop a classifier, we must first generate an ARFF file and preprocess the data.

3.4.1 Creating the ARFF file

The excel sheet containing the annotations and the corresponding text files of the articles must be combined into an ARFF file so that WEKA can access them. An ARFF file is defined as “a list of instances sharing a set of attributes” (Payton). In our ARFF file, each article is an instance and the words associated with the article are the attributes. The Apache Maven software was used to generate the ARFF file. (Appendix F: Downloading WEKA and Generating an ARFF File demonstrates how to use Apache Maven to create one’s own ARFF file.)

3.4.2 Preprocessing and Filtering the Data

After opening the ARFF file in WEKA, all the words associated with the articles are stored in the attribute, “text.” We used the string-to-word-vector filter to access the individual word vectors from the “text” attribute and to customize the words to our liking. The word vectors are then added to WEKA’s word dictionary and to be used to construct the decision tree. The string-to-word-vector filter provides multiple customization options including “lowerCaseTokens,” “minTermFreq,” “outputWordCounts,” “stemmer,” “tokenizer,” “useStoplist,” and “wordstokeep.” Figure 2 below shows the previously mentioned parameters as they appear on the string-to-word-vector menu.

Figure 2. String-to-word-vector menu with parameters of interest highlighted in red

weka.filters.unsupervised.attribute.StringToWordVector

About

Converts String attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings.

More

Capabilities

IDFTransform False

TFTransform False

attributeIndices first-last

attributeNamePrefix

doNotOperateOnPerClassBasis False

invertSelection False

lowerCaseTokens True

minTermFreq 1

normalizeDocLength No normalization

outputWordCounts False

periodicPruning -1.0

stemmer Choose **NullStemmer**

stopwords Weka-3-6

tokenizer Choose **WordTokenizer -delimiters " {\r\n\t.,;:\'}"**

useStoplist True

wordsToKeep 1000

3.4.2.1 WEKA Customization Options

LowerCaseTokens: This parameter determines whether the filter ignores the distinction between capital and lower case letters when adding words to the word dictionary. For example, when this parameter is set to the default value of false, the words “Revolution” and “revolution”

are considered as two separate attributes. For our experiment, we set this parameter to true because we found that ignoring capital letters helps increase a classifier's classification accuracy.

MinTermFreq: This parameter determines the minimum frequency of a word per class needed for the filter to add the word to the word dictionary. We kept the default value of 1 since we wished to keep all possible words.

OutputWordCounts: This parameter determines whether the words in the word dictionary do or do not include their corresponding frequencies in each article. In other words, this parameter provides two options when using a classifier: to either simply consider the presence or absence of a word when constructing a decision tree, or to also take into account how many times a word appears per article. According to our annotation process, only a single relevant key word was needed to answer yes to its corresponding question. For example, an article only needed to mention "socialism" once for the political annotation question to be answered in the affirmative. Thus, we set this parameter to the default false value because it more accurately reflected our annotation process.

Stemmer: This parameter allows the filter to include the stem of the word and not the entire word in its word dictionary. For example, when the "NullStemmer" is selected (i.e. no stemmer ability is applied), the word dictionary will include "art," "artist" "artistically," "artists," and "arts." Once a stemmer is applied, such as the "LovinsStemmer," only the word "art" appears. We found that the stemmer's impact on classification accuracy varies for each annotation question. Also, based on previous research, stemmers may help improve classification accuracy (Sanderson and Watry 78). Thus, we allowed the classifier to construct decision trees twice for each annotation question: with and without a stemmer.

UseStoplist: When this parameter is marked as true, the filter does not include a predetermined set of words in its word dictionary. These words, such as “a,” “is,” “of,” and other articles, to be words, and prepositions will not help in determining an article’s class and in fact might actually hinder the process by being included in the decision tree. Therefore, we used WEKA’s default stop word list to help improve its classification accuracy.

Tokenizer: This parameter determines the type of n-grams that the filter places into the word dictionary. As mentioned in the literature review, previous studies contradict each other on whether n-grams help improve classification accuracy (Peng and Shuurmans 14; Bekkerman and Allan 7). For our dataset, we found that using bigrams contradicted our use of the stop word list. Unigrams that were removed such as “and,” “is,” and “the,” reappeared as bigrams in “and a,” “is in,” and “the man” to name but a few possibilities. Due to the great variety in possible bigrams, it is extremely difficult to create an adequate stop word list to remove all the unnecessary and meaningless bigrams. Thus, we decided to remove bigrams and solely use unigrams when creating our decision trees.

Wordstokeep: This feature tells the filter how many words from our articles to keep in its word dictionary. Since determining the best parameters to use in the string-to-word-vector filter is largely a trial and error process, we decided to create two sets of decision trees: one in which 1000 words are kept, and in the other 10,000 words are used. Due to the great size of our dataset, we felt that only using 1000 words in the word dictionary (i.e. the default setting for this parameter) may not adequately represent the variety of text in the dataset.

Other Parameters on the String-to-word-vector Menu: The remaining parameters found in the string-to-word-vector filter’s menu do not apply to our dataset and were left as their default settings. These parameters and their default settings can be found in table 1 below.

Table 1. Default settings of unused parameters found in the string-to-word-vector filter

Parameter	Setting
IDF Transformation	False
TF Transformation	False
Attribute Indices	First-last
Do Not Operate on Per Class Basis	False
Invert Selection	False
Normalize Document Length	No normalization
Periodic Pruning	-1.0

3.4.3 The Four Final String-to-Word-Vector Filter Configurations

Overall, we used WEKA to construct four decision trees for each annotation question so we could find which set of parameters maximizes our classification accuracy. For the first decision tree, a null stemmer was used, and 1000 words were kept in the word dictionary. For the second decision tree, a null stemmer was used, and 10,000 words were kept in the word dictionary. For the third decision tree, a stemmer was used, and 1000 words were kept in the word dictionary. For the fourth decision tree, a stemmer was used, and 10,000 words were kept in the word dictionary. A combination of parameters is known as a configuration. Thus, these four configurations allowed us to find the best decision tree with the highest classification accuracy and ROC area. Table 2 below shows the four possible configurations.

Table 2. The Four Configurations used by the String-to-word-vector filter to develop word dictionaries

Configuration Number	Lower Case Tokens	Min Term Freq	Output Word Counts	Use Stoplist	Tokenizer	Stemmer	Words to Keep
1	True	1	False	True	WordTokenizer (i.e. only unigrams)	<i>Null Stemmer</i>	<i>1000</i>
2	True	1	False	True	WordTokenizer (i.e. only unigrams)	<i>Null Stemmer</i>	<i>10,000</i>
3	True	1	False	True	WordTokenizer (i.e. only unigrams)	<i>Lovins Stemmer</i>	<i>1000</i>
4	True	1	False	True	WordTokenizer (i.e. only unigrams)	<i>Lovins Stemmer</i>	<i>10,000</i>

Note that only the last two parameters (stemmer and words to keep) change for each string-to-word-vector filter configuration. All other parameters are held constant.

(Appendix G: Preprocessing the Data and Using Machine Learning Algorithms for a complete guide and instructions on how to successfully upload an ARFF file to WEKA and how to process the text data using the different parameters found in the string-to-word-vector filter.)

3.4.4 J48 Decision Tree Classifier and the ZeroR Classifier

Once our text attribute is filtered into its individual token components, we use WEKA to construct four decision trees per each annotation question. In addition to a decision tree, WEKA provides the classification accuracy rate and ROC area that allows us to assess the validity of the trees.

We selected the “J48 decision tree” option from the classification panel since it is the most popular and interpretable classification algorithm in text mining. The J48 classifier’s accuracy rate and ROC area were then compared to those of the ZeroR classifier, which can also

be found in the classification panel. The ZeroR classifier provides the baseline accuracy rate by predicting the majority class, and the baseline ROC area of approximately 0.5. Thus, we used the results of the ZeroR classifier for each annotation question to compare and assess validity of the corresponding results for each J48 classifier. A J48 classifier with an accuracy rate and ROC area greater than the corresponding ZeroR values for the same annotation question is one that is statistically superior to a model that simply picks the majority class.

3.4.5 Tenfold Cross Validation

Since we have annotated 241 articles, the classifier knows the answer to the annotation questions for each of these documents. These 241 annotated articles are then divided into training and test sets. In order to acquire an accuracy rate and ROC area, the decision tree is first developed on a training set and then applied to a test set. Thus, the classifier can use these answers (i.e. the class of each instance) and the corresponding words associated with each instance (i.e. the attributes) to develop a decision tree based on the training set using the divide-and-conquer approach. Then, this tree is applied to the test set and each instance is classified as a specific class. Since, WEKA has access to the correct classes for each instance, an accuracy rate and ROC value can be calculated. But the question remains on the best method to divide the 241 annotated articles into training and test sets. The most viable answer is to use tenfold cross validation (Witten, Frank, and Hall 153).

In tenfold cross validation, the dataset is randomly divided into ten parts of approximately the same size. The classifier uses nine out of these ten parts as a training set. Once a model is developed, it is tested on the remaining part. This process is repeated ten times in which a different part is left out for testing each time. The accuracy rate for each of the ten trials is then averaged to calculate the accuracy rate that WEKA then displays (Witten, Frank, and Hall

153). (Appendix G: Preprocessing the Data and Using Machine Learning Algorithms also includes detailed instructions on how to run a J48 decision tree and ZeroR classifier in WEKA for those unfamiliar with the program.)

3.5 Topic Modeling

Furthering our goal of discovering what kinds of technology we can use to analyze qualitative data and how we can utilize these technologies, topic modeling was suggested by our MITH advisor, Travis Brown, as a viable tool to gather information on our dataset. Specifically, this would classify our database into a number of topics that would essentially represent the different themes our entire database had. Each topic would hold a list of words that the software considered important for determining what the topic would be. Furthermore, these lists of words would be listed from most important to least important to the topic.

Figure 3. Example of a topic generated from our database of articles

topic-27	god	religion	church	christ	christian	faith
----------	-----	----------	--------	--------	-----------	-------

Figure 3 is a strong religion topic. The “topic-27” is an arbitrary label for the topic while the words “god,” “religion,” “church,” “Christ,” “Christian,” and “faith” make up the topic/theme. The word “god” would have a heavier weight in the topic than the word “religion,” as “religion” would be weighted more than “church” and so on. These word lists would span many columns. Ultimately, our goal was to take each topic and plot it against time to see how prevalent a topic was thought the years between 1890 and 1945.

3.5.1 Creation of the Topics Spreadsheet

After generating our database of articles via scanning and OCR, we put all of the text of every article into a single TXT file that would be used by the program, MALLET to generate a MODEL file. Through the use of MALLET, we were also able to manipulate variables such as

the number of total topics generated and the stop words list – a list of words that we consider irrelevant, noise, or nonsense. However, as this was our first use of MALLET, we did not alter anything of the variables and used the presets provided to us by Travis Brown. After generating a MODEL file, we then used the Apache Maven software again to translate this MODEL file into a more easily readable EXCL file of topics.

3.5.2 Adjustment of Topics

One of the first things we noticed about our topics was that there were a lot of irrelevant or nonsense words such as “thou”, “thee”, “thy”, “ame”, “iie”, and “iii” included in the topics. More alarming was the fact that these words appeared in the first 20 words of each topic, which constituted the most important words in determining the topic. Thus, we decided to “clean up” our topics by adding these words to the stop word list and regenerating our MODEL file. Ultimately, we ended up with stronger and clearer topics after this process.

Another variable we manipulated was the number of topics we could generate. We found that as we decreased the number of topics, the topics became more general. Please see Figure 4.

Figure 4. Selection of topics with the number of topics set to 10

topic-00	russian	great	book	literature	work	author
topic-01	russia	russian	people	governme	years	country
topic-02	pushkin	poet	russian	pierre	onegin	poem
topic-03	life	man	russian	gorky	world	love
topic-04	tolstoy	man	life	men	people	world
topic-05	turgenev	madame	paris	gorski	vera	love
topic-06	soviet	russia	russian	revolution	literature	war
topic-07	tolstoy	time	father	life	count	family
topic-08	man	eyes	people	time	long	day
topic-09	play	chekhov	theatre	plays	art	stage

No topics relating to religion even exist with only ten topics, and the topics were relatively general. On the other hand, we found that increasing the topics made them more specific only up to a point— any additional topics generated make the lists once again unclear and increase repetitiveness, as evidenced in figure 5.

Figure 5. Selection of topics with the number of topics set to 80

topic-31	gorki	maxim	novgorod	capri	alexei	khan
topic-38	gorky	maxim	life	social	tramp	tramps

With 80 as our number of topics, there were two “Gorky” topics that are inexplicably separated. Many more examples of this sort of ambiguity throughout our 80-topic spreadsheet existed. Ultimately, through a process of trial and error, as well as consulting with our MITH advisor, Travis Brown, we determined that 40 created the most clear and encompassing topics.

3.6 Regression Methodology

Although topic modeling alone provides valuable information in terms of how words are grouped together, its application needs to be expanded to answer more specific questions a researcher may want to pose. We created three distinct experiments to test how we can more usefully apply topic modeling output.

Experiment 1: Test of time variant data (sea change). This is the ability to distinguish between articles pre- and post- a certain time and it is the simplest experiment type.

Experiment 2: Test of time invariant data. Answering some of the annotation questions is our method of running this experiment since it involves differentiating articles on the basis of a non-time determined attribute such as the name of the primary Russian author discussed.

Experiment 3: Test of quasi time variant data. This involves determining whether articles belong to certain kinds of time periods such as those of war and crisis. These time intervals are unlike

the ones of the sea change because they are not exclusively pre- or post- a specific date, increasing the difficulty of the experiment.

In order to do conduct these experiments, we need to be able to create statistics based models that classify articles objectively. Since there is a single desired output in these experiments, the placement of an article into one of two categories, and multiple inputs, a regression model is suitable.

3.6.1 Factor Analysis

In big data regression modeling applications, the number of variables captured is often excessive. There is a strict need to reduce the volume of variables while still capturing the individual significance of all of the variables. For purposes of creating a regression model, having too many variables can lead to error and insignificant variable coefficients because of collinearity among some of the variables (Bai and Ng 91). Our dataset for topic modeling is made up of 40 topic variables. Although, this is number of variables is not as large as those produced from mining other big data sources, ranging in the hundreds, decreasing this number would still be useful for improved regression results.

Factor analysis captures how all of the variables move relative to one another across the dataset and produce factors to compound similar movements (Bai and Ng 93). Specifically, common factor analysis attempts to generate the minimum number of factors necessary to describe the correlation among variables given the percentage value of how much of the correlation is captured by the factors. Using factor analysis compatible statistics software, SPSS in our case, the 40 topics were narrowed down to 24 factors for use in creating a regression model. The component matrix showing the makeup of each factor is shown in Appendix H: Partial Rotated Component Matrix for Factor Analysis Data (Factors as Columns).

3.6.2 Stepwise Binary Logistic Regression

Stepwise Binary Logistic Regression (SBLR) is a regression method in which the explained variable Y is categorical (binary) and all explanatory variables that are initially inputted into the regression are only included if they are significant in determining Y . The logistic aspect specifies that the explained variable is binomial (categorical) as opposed to a linear regression model in which the explained variable is continuous. The regression methodology selects variables one by one in terms of the most significant variables in predicting Y . Therefore, the resulting list of variables and their coefficients is sorted by the order of importance in predicting Y . An additional advantage of SBLR is that assumptions of normality and homoscedacity are not necessary for the variables. Proving normality and homoscedacity for topics and factors would otherwise prove to be very difficult.

3.6.3 Topic Modeling Experiment 1

The Bolshevik Revolution is of key interest given that it is a crucial turning point in Russian history. It is valuable to see if the data produced by topic modeling and refined by factor analysis is able to create a SBLR model that predicts whether an article was written before or after a sea change event, such as the Bolshevik Revolution, with significant accuracy. This reveals which U.S. media topics, and therefore the corresponding ideas, that the Bolshevik Revolution had influenced.

Our hypotheses for this experiment are as follows:

H_0 : No difference pre- and post- Bolshevik Revolution (Result: Low accuracy)

H_a : Difference pre- and post- (Result: High Accuracy)

Before constructing this experiment, it is important to note that although the majority of the Bolshevik revolution occurred in 1917, a reporting delay is conjectured given the distance

between Russia and the US, and the speed at which the information would surface in the US discourse. This gap could be one year or a few years, making it necessary to run SBLR for different years and updating the hypotheses to read in terms of pre- and post- a hypothesized sea change year. This sea change year is theorized to be the year in which a significant change occurred in the topics discussed in US commentary on Russian Literature, as a result of the Bolshevik Revolution.

Using SBLR, we included both topics and factors, contrary to the variable minimizing purpose of the factors. This was done because although the factors would eliminate the need for topics, inputting individual topics would provide a stronger narrative for describing the shift from before and after the sea change year. We initially expected that the factors would contribute to the model by improving the significance of the model's coefficients.

3.6.4 Topic Modeling Experiment 2

Beyond supporting existing frameworks and theories regarding American reception of Russian literature and Russia in general, our work in annotation allows us to examine the power of SLBR and linear regressions to predict primary authors and a number of other annotation questions. We construct both SLBR and linear probability regressions to determine improvement on base accuracy from both models. In both cases, we input every article's factor data derived from topic modeling into the appropriate model to predict annotation responses. This allowed us to test the ability of our Topics Modeling data to predict responses to time independent queries.

3.6.5 Topic Modeling Experiment 3

It may be of interest to explore alternative hypothesis regarding predicting political events beyond the fundamental sea change event idea explored in Experiment 1. We consider

predicting whether articles were written during years of foreign or domestic crisis are differentiable with an SLBR model from articles written outside of these years.

Our hypotheses for this experiment are as follows:

H₀: No difference between Russian crisis and non-crisis years

(Result: Low accuracy)

H_a: Difference between Russian crisis and non-crisis years

(Result: High Accuracy)

As in Experiment 1, we must recognize there may be lags in the dissemination of information to the United States as well as a delay in the creation of an potentially unobserved consensus developing in the United States, which could drive a change in language and in turn create differences in language between articles written during the crisis period and not.

Chapter 4: Results and Discussion

4.1 WEKA

4.1.1 General Trends

In this study all ZeroR accuracy rates, J48 accuracy rates, and J48 ROC areas were calculated by WEKA. The most successful configuration for each annotation question was determined by elevating the J48 accuracy rate. The highest accuracy rate for each question was considered the most successful configuration since our goal was to find the classifier with the most accurate prediction rates. In the two cases in which two configurations produced the same accuracy rate, the ROC area was used to distinguish between the two. Since the ROC area is considered a measure of a classifier's predictive capabilities, it is an appropriate measure to use to choose between two classifiers with equal accuracy rates. Table 3 lists the most and least successful configurations along with the accompanying J48 accuracy rates and ROC areas for each of the annotation questions not pertaining to sentiment. Table 4 lists the same information for the annotation questions relating to sentiment analysis. The ZeroR accuracy rate is provided as a baseline for comparison. By definition, the ZeroR classifier has an ROC area of approximately 0.5. Thus, a classifier with a J48 accuracy rate and J48 ROC area greater than the corresponding ZeroR values indicates that the model is statistically superior to simply picking the majority class for each instance.

Table 3. The most and least successful configurations for each annotation question with their respective accuracy rates and ROC areas

<i>Annotation Question</i>	<i>ZeroR Accuracy Rate (%)</i>	<i>Most Successful Configuration</i>	<i>J48 Accuracy Rate (%)</i>	<i>J48 ROC Area</i>	<i>Least Successful Configuration</i>	<i>J48 Accuracy Rate (%)</i>	<i>J48 ROC Area</i>
Principal Author Subject of Debate	73.0290	4	74.2739	0.651	3	59.7510	0.494
Books Mentioned	79.668	1	70.1245	0.598	3	66.3900	0.555
National Identification	64.7303	1	71.3693	0.698	3	68.0498	0.685
Style	49.7925	2	65.9751	0.659	1	59.3361	0.585
Gender	65.1452	3	67.2199	0.648	1	59.3361	0.0715
Race	70.5394	2	74.2739	0.690	1	64.7303	0.572
Socioeconomic Class	55.6017	2	67.6349	0.662	4	59.3361	0.565
Religion	56.4315	3	67.6349	0.638	1	60.9959	0.603
Radical Politics	65.5602	4	75.1037	0.741	1	70.5394	0.643
Russia as topic of Similarity to the West	66.3900	4	71.7842	0.639	2	58.0620	0.505
Russia as a topic of Contrast to the West	64.7303	1	71.3693	0.654	2	61.8257	0.577
Foreign Place Names	78.8382	2	73.8589	0.606	1	67.2199	0.557
Gender of Article Writer	85.9375	1	79.6875	0.509	3	71.8750	0.482

Table 4. The most and least successful configurations for each sentiment analysis question with their respective accuracy rates and ROC areas

<i>Sentiment Analysis</i>	<i>ZeroR Accuracy Rate (%)</i>	<i>Most Successful Configuration</i>	<i>J48 Accuracy Rate (%)</i>	<i>J48 ROC Area</i>	<i>Least Successful Configuration</i>	<i>J48 Accuracy Rate (%)</i>	<i>J48 ROC Area</i>
General Opinion: Positive vs. Negative vs. Neutral/ Mixed Opinion	53.5270	2	57.6763	0.591	3	46.8880	0.483
Positive vs. Negative	91.9643	2	89.2857	0.502	3	83.9286	0.451
Positive vs. Non-positive	57.2614	2	60.9959	0.588	1	50.6224	0.468

For the text classification data not pertaining to sentiment, no trend emerged in which one configuration consistently outperformed the rest. In fact, configuration, 1, 2, and 4, were ranked as the most successful configuration for 3 annotation questions each, and configuration 3 was assessed as the most successful for 2 annotation questions. Similarly, no configuration consistently underperformed when compared to the others. All configurations appeared at least once in the least successful configuration category with configuration 1 appearing the most for a total of six times. For the sentiment analysis data, configuration 2 was ranked the best suggesting that sentiment analysis questions require access to a large amount of vocabulary to construct the most accurate tree. However, the same lack of consistency was observed in the sentiment analysis data in that no single configuration was the least successful. This variability in results confirms one of the downsides of using WEKA to analyze large amounts of data when compared to topic modeling: WEKA requires significantly more human input than topic modeling without producing more conclusive results. Not only did the classifying algorithms require us to annotate hundreds of articles, each specific annotation questions has its own word dictionary

configuration that maximizes its classifier's accuracy. From our results, we cannot conclude that any one specific configuration produces the best word dictionary that in turn results in the best J48 decision tree. The topic modeling data will be discussed later in this section. (Appendix I: Results of All Four Decision Tree Configurations for Each Annotation Topic for a detailed look at the accuracy rates and ROC areas for all 4 configurations for each annotation question).

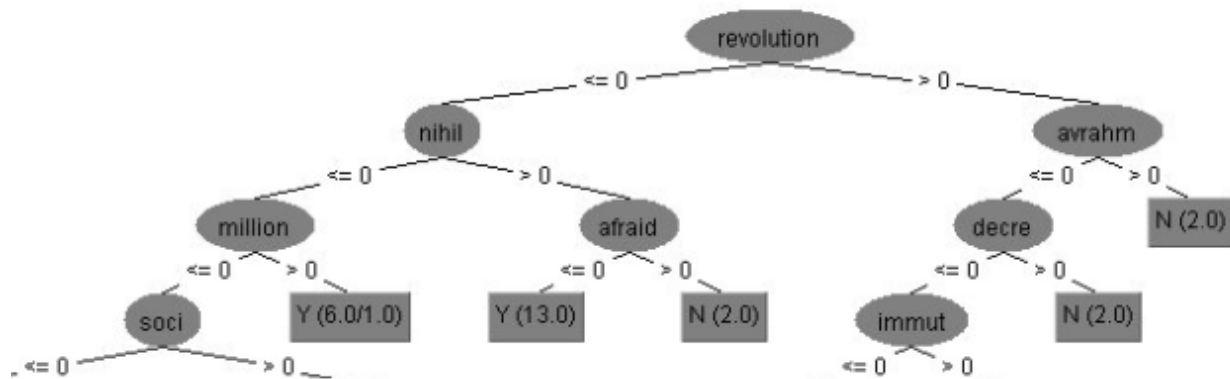
All of our classifiers fell short of achieving an extremely high classification accuracy rate approaching 90%. However, our accuracy rates are similar to those found in some published research (Lee et al. 256). Furthermore, all our classifiers were still statistically significant and contain predictive qualities as is supported by ROC areas greater than 0.5. Recall that a ROC area of approximately 0.5 signifies a classifier with no predicative qualities and one that simply predicts the majority class. We discuss how to improve the accuracy rate of our classifiers by the end of this section, but for now, we will consider the specific J48 classifiers for each text classification annotation question by dividing them into five categories: 1) the most successful classifiers for this dataset; 2) the classifiers associated with annotation questions that were difficult to annotate for; the 3) the classifiers that experienced a modest improvement in accuracy when compared to the ZeroR accuracy; 4) the classifiers that had a worse accuracy rate than that of the ZeroR classifier; 5) the classifiers of the sentimental analysis annotation questions. For convenience, we present only segments of each decision tree in the discussion below.

4.1.2 The Most Successful Classifiers

The classifier for the radical politics question achieved the greatest accuracy rate and ROC area for the entire dataset. Thus, this is our most successful classifier. Among the key political words that we searched for as we annotated the text were “revolution,” “nihilism,” and “socialism.” The J48 decision tree also used these words in developing its model as seen in

figure 6. In fact, the decision tree has an extremely methodological approach because it first looks for “revolution,” a political word often associated with Russia, in the text. If this term is not found, it then goes to other political thoughts often surfacing in Russian literature and history such as nihilism and socialism. Note, the stems “nihil” and “soci” are used instead of “nihilism” and “socialism,” respectively.

Figure 6. A segment of the J48 decision tree for the annotation question pertaining to radical politics

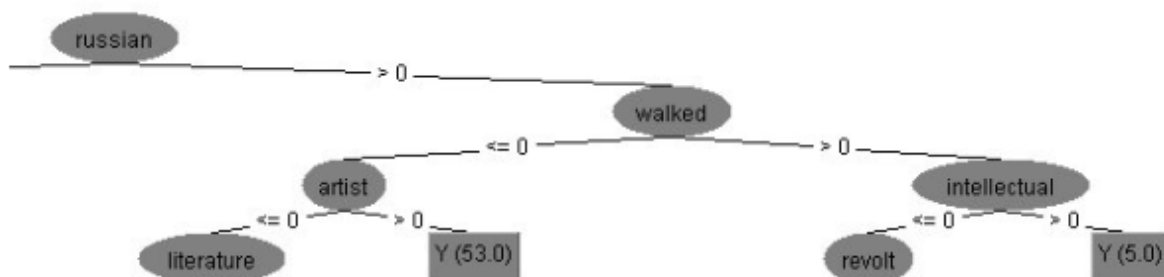


Note that the numbers in parentheses, with the form (A/B), relate to the training set and not the test set. The A is the number of instances that reach that tree in the training set, and the B is the number of incorrectly classified instances. If no B value is used then no instances were incorrectly classified. Thus, “Y(6.0/1.0)” from the decision tree above signifies that six instances reached this node, 1 of which was incorrectly misclassified. The remaining five instances were correctly classified as “Y” (i.e. yes).

The classifier for the national identification annotation question had an accuracy greater than 70% and one that increased by 6.64 percentage points when compared to the ZeroR accuracy. Also, the decision tree accompanying this model correctly reflects the methodology used by our annotators to answer this question. The annotators looked for the text to specifically

mention that a literary author was Russian or draw the reader’s attention to the Russianness of his works. Thus, the fact that the very first word on the partial decision tree is “Russian” as shown in figure 7 signifies that this model partially matched our annotation process. However, unlike in our annotation process in which the mention of the word “Russian” with an author was enough reason to classify this question as a yes, the model went a few steps further. It also included the absence or presence of words such as “artist” and “intellectual” as part of the decision process. This decision tree thus reveals that Russianness is generally associated with specific characteristics such as “artist” and “intellectual.” In fact, for every instance, the classifier found that documents in our training set that included “Russian” and one of these words were annotated as invoking the nationality of the author. This in itself is an interesting finding. Russian literary authors could have been associated with much more neutral terms such as novelists, poets, and writers. But instead we find them associated with the more substantial and thought provoking terms, “artist” and “intellectual.”

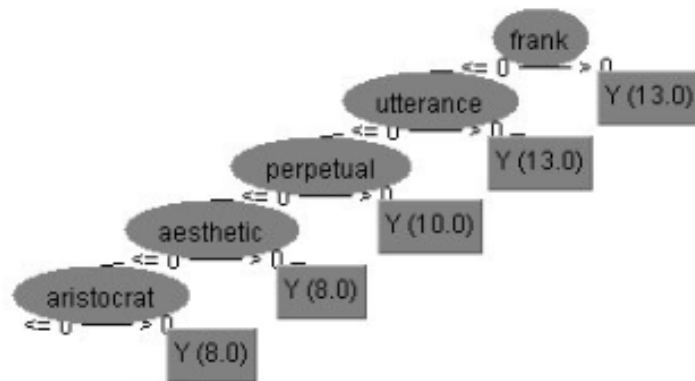
Figure 7. A segment of the J48 decision tree for the annotation question pertaining to the author’s national identification



The style annotation questions experienced the greatest increase in accuracy when comparing the J48 classifier to the ZeroR classifier. In this case, the increase was 16.1826 percentage points. Since the ROC area was 0.659, this model has predictive abilities. For this

question, the annotators looked for the text to describe a Russian author or piece of literature in terms of stylistic and artistic capabilities. In the partial decision tree shown in figure 8, only the fourth word, “aesthetic,” directly refers to a stylistic property. “Frank” and “utterance” may be terms used by an article writer when discussing an author’s writing style. But more importantly, the inclusion of “perpetual” and “aristocrat” in the decision tree illustrates that the style topic is concerned with many other non-literary factors. In this case, style is associated with discussions of social class (i.e. “aristocratic”) and history (i.e. “perpetual”). In other words, this decision tree suggests that literary aspects of a novel are inseparable from discussions of historical and social phenomena of the time or those found in a specific text.

Figure 8. A segment of the J48 decision tree for the annotation question pertaining to the style and artistry of the principal author



Even though both the socioeconomic class and religion annotation question had an accuracy rate below 70%, they both experienced some of the greatest increases in accuracy found in our dataset with a 12.0032 percentage points and 11.2034 percentage points increase, respectively. The J48 decision tree produced from the socioeconomic class question in figure 9 also includes many words that one would associate with socioeconomic class such as “serfs,” “nobility,” and “aristocracy.” Likewise, the decision tree associated with the religion annotation

question found in figure 10 includes all words derived from “Christ” and “religion.” Furthermore, this decision tree illustrates some of the difficulties associated with using a stemmer. For example, the very first stem used by the classifier is “chr.” We are unsure which word this stem originates from and WEKA does not provide us with any additional information. Also, it is possible that the word “Christ” was erroneously stemmed into “chr.”

Figure 9. A segment of the J48 decision tree for the annotation question pertaining to socioeconomic class

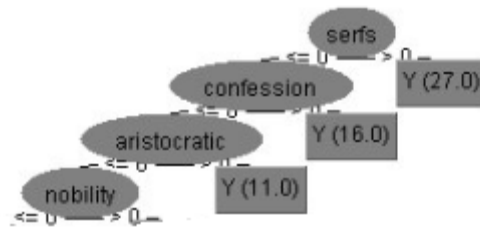
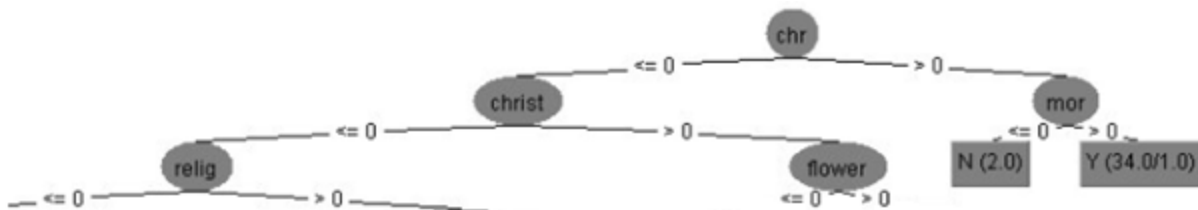


Figure 10. A segment of the J48 decision tree for the annotation question pertaining to religion



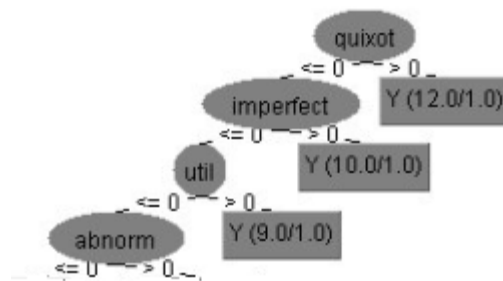
4.1.3 Classifiers for Questions Difficult to Annotate

The annotation questions concerning the similarities between Russia and the West, the differences between Russia and the West, and whether the principal author is a subject of debate were among the most difficult to annotate for. Unlike the other text classification questions, these questions lacked specific key words to look for. Furthermore, in comparing Russia and the West or in debating a Russian author, an article writer is likely to use a great deal of nuances that we felt would be missed by a classifier. Thus, we were pleasantly surprised that all 3 classifiers had

an accuracy rate above 70% and at the relative modest increase in accuracy when compared to that of the ZeroR classifier.

The J48 classifier for the annotation question concerning the similarities between Russia and the West had a modest improvement of 5.3942 percentage points when compared to the ZeroR’s baseline value. As the partial decision tree shows in figure 11, “quixot” is the first word used by the decision tree. “Quixot” is likely a stem for the adjective “quixotic” which in turn refers to *Don Quixote*, one of the classics of modern Western literature. Thus, it is possible that in describing the recent phenomena of Russian literature, many article authors from our dataset used *Don Quixote* as a comparison.

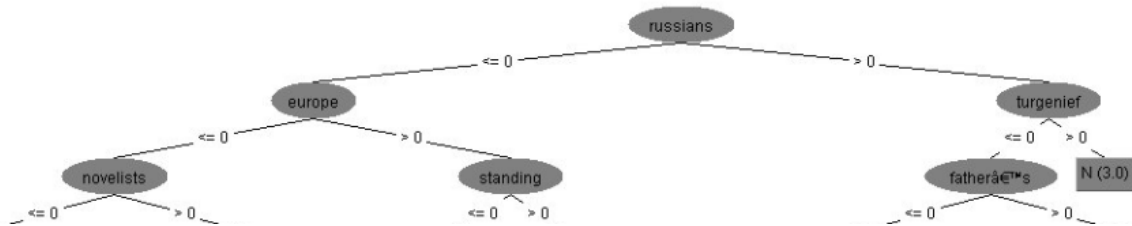
Figure 11. A segment of the J48 decision tree for the annotation question pertaining to the similarities between Russia and the West



The classifier for the annotation question contrasting the Russia and the West had a modest improvement of 6.6390 percentage points when compared to the ZeroR’s accuracy rate. From reading the decision tree in Figure 12, we see that if “Turgenief” is mentioned in an article, then the article is not contrasting Russia and the West. This conclusion made by the decision tree reflects literary scholars’ understanding of Turgenev since he is considered to be a more Western and French-like than most other Russian authors. The word, “Europe,” also seems appropriate

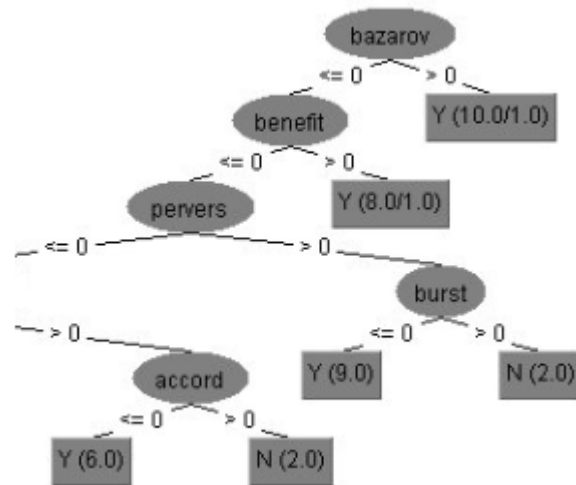
since it might reference the dichotomy that is still discussed today of whether Russia is a European nation or something on to its own.

Figure 12. A segment of the J48 decision tree for the annotation question pertaining to the contrast between Russia and the West



The J48 accuracy rate was only 1.2449 percentage points higher than the ZeroR accuracy rate in determining whether the principal author is a subject of debate. However, the ROC area of 0.651 is greater than 0.5, which signifies that this model does have significant predictive qualities. Unlike the previous two nuance-filled questions where we can get a general idea of the “thinking” behind the decision tree, the partial decision tree found in figure 13 left us with many more questions. “Bazarov” refers to the main character in Turgenev’s *Father and Son*. This controversial nihilist character may have been used in debating the merits of Turgenev’s works. The other word vectors, in this case stems of words, have no apparent connection to words that may be used in a literary debate. Furthermore, as table 3 shows, the worst configuration for this annotation question had an accuracy rate of 59.751%, 13.2780 percentage points below the baseline value. The great deal of nuance used by article writers when debating a literary author and the corresponding lack of key words, are the most likely explanation to why WEKA struggled to produce a classifier with a greater increase in accuracy rate for this question.

Figure 13. A segment of the J48 decision tree for the annotation question pertaining to whether the principal author is a subject of debate



4.1.4 Classifiers with a Modest Improvement in Accuracy

Unlike our most successful classifiers that experienced gains in accuracy of over 9%, the classifiers for the gender and race annotation questions had more modest results. Even though only a small gain of 3.7345% is observed when using the J48 classifier for the race annotation question, the decision tree in figure 14 uses race related terms that our annotation team searched for such as “slav” and “negro.” Interestingly, the words “surveillance” and “alarmed” are associated with race. This suggests that racial topics are associated with an anxious, apprehensive, and possibly fearful tone that reflects common held viewpoints of the time period. Similarly, for the gender question, a modest increase of 2.07 percentage points was observed. Once again, we see that the decision tree in figure 15 reflected our annotators’ thinking process since words like “woman” were used in the classifier.

Figure 14. A segment of the J48 decision tree for the annotation question pertaining to race

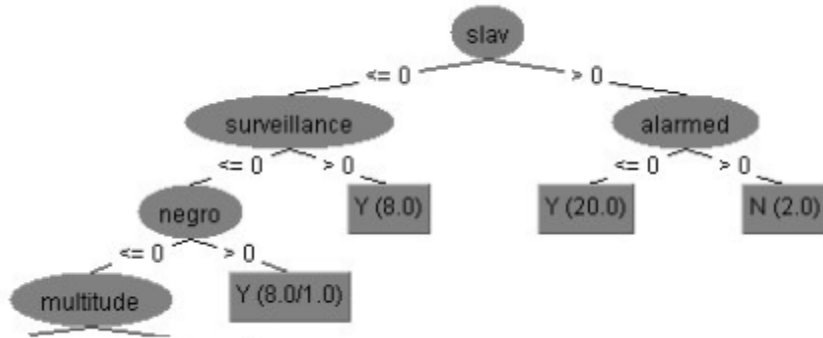
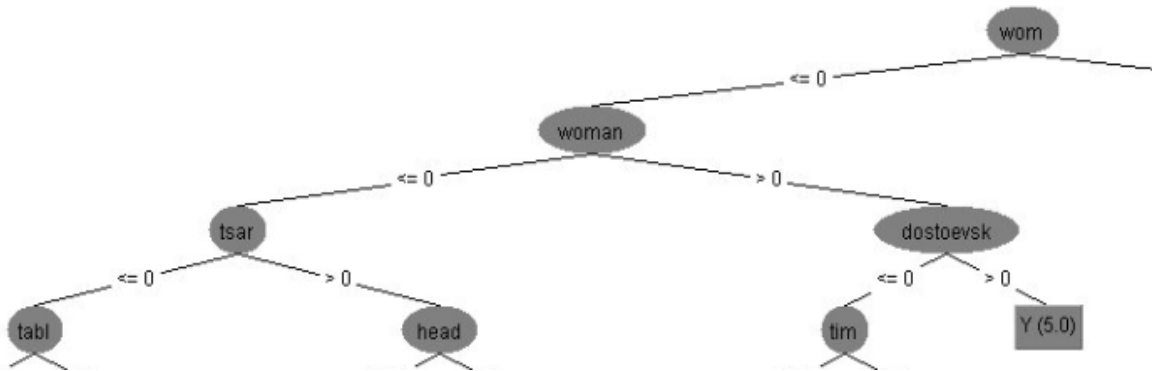


Figure 15. A segment of the J48 decision tree for the annotation question pertaining to gender



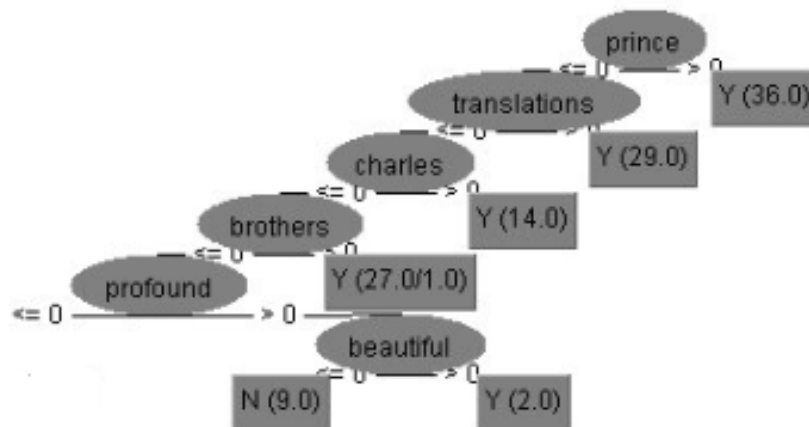
4.1.5 Underperforming J48 Classifiers

In the text classification experiments, three J48 models performed worse than the ZeroR classifier: the books mentioned, foreign place names, and gender of the article writer annotation questions. The best J48 configuration for the books mentioned annotation question was 70.1245%, which is 9.5435 percentage points lower than the ZeroR accuracy of 79.668%. Similarly, best J48 classifier for the foreign place names question only musters an accuracy rate of 73.8589%, which is 4.9793 percentage points below the ZeroR accuracy of 78.8382%.

However, since both these J48 classifiers an ROC area greater than 0.5, WEKA has concluded that these classifiers have predicative capabilities. Ironically, for these two categories, simply picking the majority class is more accurate than using a predictive classifier. We were surprised at the accuracy rates for these two questions since they were among the easiest to annotate for. The annotators simply looked for literary titles, such as “*War and Peace*” and “*The Demons,*” or foreign place names such “St. Petersburg” and “Yasnaya Polyana,” Tolstoy’s estate. Fortunately, the decision trees for both these categories provide insight on why these models fared poorly.

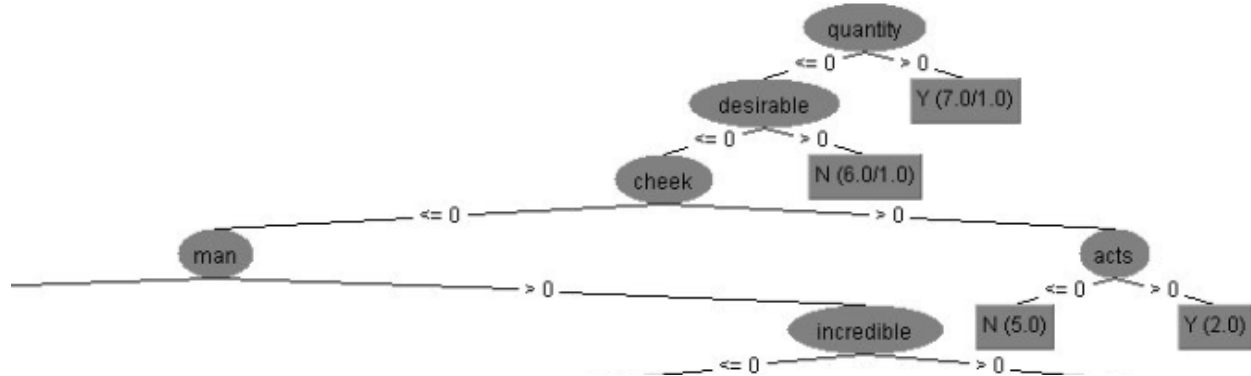
In the decision tree for the books mentioned question found in figure 16, the first word that this model considers is “prince,” followed by the words “translations,” “Charles,” “brothers,” and “profound.” The only word among these that may refer to a title of a Russian literary work is “brothers” which most likely comes from *Brothers Karamazov*. “Prince” and “Charles” might refer to names of characters, or those of actual people. When we annotated the documents for this question, we specifically looked for titles of Russian literature. The fact that the classifier looked to other characteristics to create its decision tree, may explain a J48 accuracy rate below that of the ZeroR classifier.

Figure 16. A segment of the J48 decision tree for the annotation question pertaining to the mention of books or other literary works



Similarly, as we annotated our dataset for the foreign place name questions, we searched for geographical locations. However, the J48 classifier completely ignored all words related to geography and instead focused on words such as “quantity,” “desirable,” “cheek,” “man,” and “acts” as the partial decision tree in figure 17 illustrates. Thus, we can draw two general conclusions. First, the decision trees did not reflect the thinking process used by our annotators. Second, since there is a great deal of variety in names of literary works and foreign locations, the same ones are unlikely to appear in most articles. Thus, as the decision tree suggests, the classifier was unable to find those unique words that appeared in the majority of documents that would guide it in answering these two annotation questions.

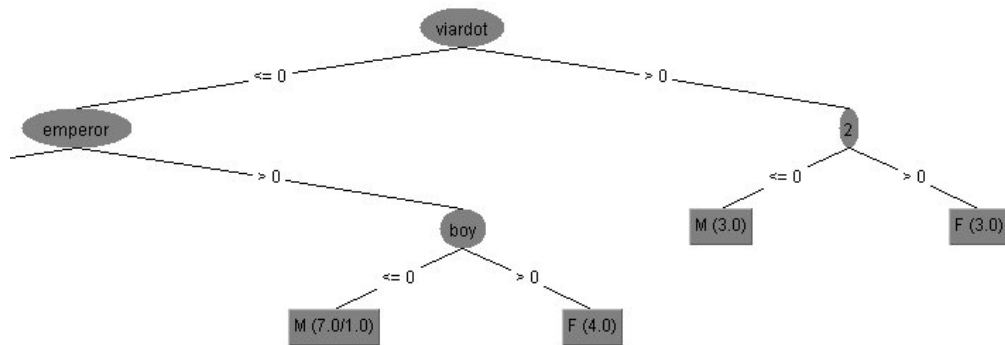
Figure 17. A segment of the J48 decision tree for the annotation question pertaining to the mention of foreign place names



The gender of the article writer annotation question required some additional editing before a classifier could be developed. Out of the 241 annotated documents, 112 were authored by men, 18 by women, and the remaining 111 lacked any claimed authorship or the gender of the author could not be determined due to the use of initials. Thus, since we wanted to develop a classifier that could distinguish between male and female article writers, we removed all the articles by unknown authors before developing the J48 classifier. A partial decision tree is

provided below in figure 18. The J48 classifier accuracy rate was 6.2500 percentage points below the ZeroR accuracy rate. Furthermore, since the ROC area is 0.509 and is barely above 0.5, this classifier has no predictive qualities. Even the partial decision tree in figure 18 does not reveal any useful information. The classifier uses the presence or absence of the numeric value “2” to classify several instances. The lack of a sufficient amount of known female article writers in the annotated dataset probably resulted in the underperforming classifier for this annotation question.

Figure 18. A segment of the J48 decision tree for the annotation question pertaining to the article writer’s gender



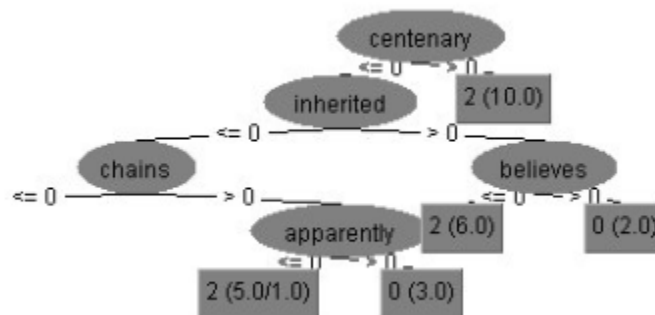
4.1.6 Sentiment Analysis

The classifier struggled to develop adequate decision trees for the sentiment analysis questions compared to the decision trees produced for the other annotation questions. These results can be explained primarily by the nature of the dataset and the use of language to convey emotion.

For the general opinion sentiment analysis question, the J48 classifier’s accuracy rate was 4.1493 percentage points more than that of the ZeroR classifier. The accuracy rate of 57.6763% is lowest accuracy rate found in our dataset. The partial decision tree in figure 19 shows that the

classifier struggled to find any keywords that denoted whether an article was of positive, negative, or a neutral/mixed opinion. This finding can be accounted for by the great variety of ways in which an article writer may praise or defame a Russian author. For example, the presence of the word “centenary” signifies that an article is marked as having a positive sentiment. This makes sense because “centenary” likely refers to the celebrations and honoring of Alexander Pushkin a hundred years after his death. However, the other words used, such as “apparently,” do not seem to be associated with any sentiment related traits. We felt that the instances classified as neutral/mixed option may be confusing the classifier since these instances will have both positive and negative sentiments expressed. Thus, we were removed all these instances and created a new J48 classifier.

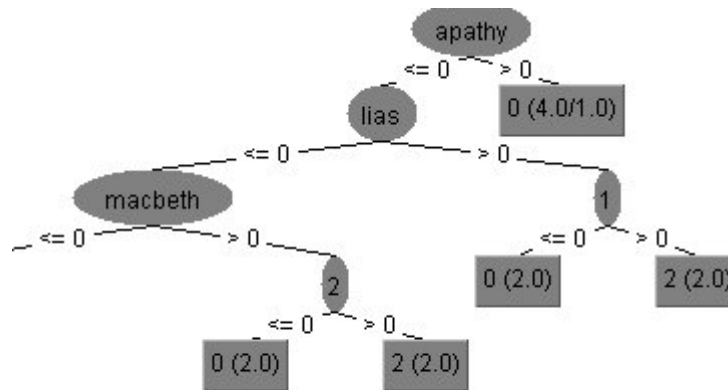
Figure 19. A segment of the J48 decision tree for the sentiment analysis annotation question for positive, negative, and neutral/mixed opinion articles



At first glance, the sentiment analysis question for positive and negative articles appears to have an extraordinarily high accuracy rate of 89.2857%. But in reality this accuracy rate is 2.6786 percentage points below that of the ZeroR classifier. In other words, picking the majority class, positive in this case, for an unknown instance will give us more accurate results than the classifier. The fact that this J48 classifier has a ROC area of 0.502, which equals the baseline area of approximately 0.5, also attests to the fact that this classifier has no predictive qualities.

The decision tree in figure 20 shows that this J48 classifier could not find words that suggest positive or negative sentiment. In fact, the J48 classifier used the numeric values “1” and “2” found in the text to make its decision. However, the fault for this sentiment analysis question does not lie in the classifier, but in the dataset.

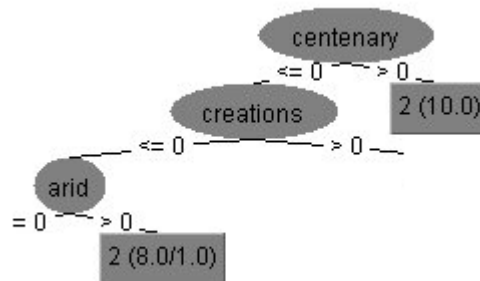
Figure 20. A segment of the J48 decision tree for the sentiment analysis annotation question for positive and negative articles



Even though the articles for the dataset were randomly selected, out of 241 articles, 103 were annotated as having a positive sentiment, 129 as having a neutral/mixed opinion sentiment, and only nine of them as having a negative sentiment. For this annotation question, since we ignored the neutral/mixed opinion questions, we were in fact applying the J48 classifier on a dataset with 103 articles with positive sentiment and 9 articles with negative sentiment. Thus, we cannot expect the classifier to find words that are characteristic of articles with negative sentiment when over 90% of the total articles fall under the positive class. Since we did not have an adequate representation of one of our classes in the training set, it is no surprise that the classifier fared poorly (Witten, Frank, and Hall 152). We attempted to mitigate this problem by developing a classifier that distinguished between positive and non-positive instances, with the non-positive class including all negative and neutral/mixed opinion articles. However, this

approach did not work since we ran into the problem that the articles annotated as non-positive, due to their inclusion of the neutral/mixed opinion articles, would also have words associated with positive sentiment. Thus, as is shown in figure 21, the J48 classifier once again used words that do not necessarily relate to sentiment such as “creations” and “arid” to classify an instance.

Figure 21. A segment of the J48 decision tree for the sentiment analysis annotation question distinguishing between positive and non-positive (i.e. negative and neutral/mixed opinion) articles



4.1.7 Suggestions for Improvement

As previously mentioned, some of our accuracy rates are similar and even slightly higher than those reported for other text datasets in the literature (Lee et al. 256). But in text classification, even though all researchers aim for a high classification accuracy rate, the creation of the dataset and preprocessing of the data are considered highly important because each text dataset differs considerably from the next (Frank, Witten, and Hall 389). Nonetheless, we propose a few suggestions that may guide any researcher who decides to work with this Russian literary dataset.

Witten, Frank, and Hall characterize text mining as an “experimental science” (403). Thus, it should be no surprise that our first general suggestion for improvement would be to try different configurations of the string-to-word-vector filter when producing word dictionaries. We

tested 4 configurations out of hundreds of possible ones. One parameter we did not experiment with was outputting words counts. Thus, the word dictionary would include the frequencies of words in addition to the word vector themselves. This parameter can also be paired with a normalization parameter in which the word frequencies would then be normalized based on document length. This would ensure that both short and long articles would have the same influence in developing the word dictionary.

Another possible suggestion would be to change the minimum word frequency to a value greater than one. Initially we kept the default value of 1 by arguing that we wanted to keep as many words as possible in the word dictionary to reflect the great variety of text found in our dataset. However, by setting the minimum word frequency value to a number greater than one, we are filtering the dataset in such a way that the word dictionary will only have often repeated words. These words are likely to be more significant than words that appear only once and thus this word dictionary may help improve a classifier's accuracy.

Even though we chose not to use bigrams and trigrams and justified our actions by stating that we did not look for a specific sequence of words when annotating, it might be worth reconsidering. As previously mentioned, some studies show significant increase in accuracy when n-grams are used (Peng and Shuurmans 14). In combination with a customized stop word list that removes all unwanted n-grams, this configuration may prove to be a powerful tool.

Nonetheless, the primary basis for creating our word dictionaries involved the bag of words theory in which text documents are considered a collection of words with word order and context being irrelevant (Nuntiyagul et al. 32-33). However, in the past decade, scholars in the field have begun to promote an updated bag of words approach that more accurately reflects the complexity of text. This new proposed method also considers the syntactic information

associated with each word. Sable et al. used an automatic part-of-speech tagger in which a computer-based algorithm was used to categorize each word in a document as a subject, verb, etc. They argued, and a portion of their results also imply, that the subject and verbs found in a text document are the most useful in text classification (176). Thus, we would also suggest to future researchers who work with our dataset to consider including a word's syntactic properties in the word dictionary.

4.1.8 Concluding Remarks

Even though the majority of our J48 classifiers outperformed the baseline values, we still felt that the accuracy rates were too low to use to predict the class of the non-annotated articles. We believe that future researchers who take into account some of the suggestions proposed above for our dataset may develop more accurate decision trees. However, even though WEKA, as a statistical tool to analyze a large dataset, did not perform to our expectations, we felt that using topic modeling provided even more information than a classifier could. Excluding the sentiment analysis information, topic modeling managed to reveal information pertaining to all our other annotation questions in much greater detail than WEKA did as is described in the next section.

4.2 Topic Modeling Experiments

4.2.1 Topic Modeling Experiment 1

Recall that we consider the creation of a model to predict whether articles were written before or after a sea change even -the Bolshevik Revolution – with significant accuracy.

As seen in figure 22, SBLR was run for all years from 1917 to 1921. The test run with 1919 as the hypothesized sea change year produced the most accurate model with an overall percentage accuracy of 77.3%. This would speculatively indicate that there is a small lag from

the beginning of the revolution (1917) to when there was a significant resulting change in the language and content of American journalistic print (1919).

Table 5. Comparison of Model Accuracy for Multiple Years and Modifications

Step 13	cutoff1917	.00	471	89	84.1
		1.00	177	393	68.9
Overall Percentage					76.5
Step 16	cutoff1918	.00	495	92	84.3
		1.00	178	365	67.2
Overall Percentage					76.1
Step 14	cutoff1919	.00	523	88	85.6
		1.00	169	350	67.4
Overall Percentage					77.3
Step 13	cutoff1920	.00	536	85	86.3
		1.00	174	335	65.8
Overall Percentage					77.1
Step 17	cutoff1921	.00	572	87	86.8
		1.00	169	302	64.1
Overall Percentage					77.3
Step 15	out1920_1922	.00	544	67	89.0
		1.00	143	282	66.4
Overall Percentage					79.7
Step 15	out1920_1923	.00	560	51	91.7
		1.00	136	232	63.0
Overall Percentage					80.9

However, with the expertise of Dr. Ronna Mallios, we discovered that a large portion of the predictive error in our model was occurring in the years 1920 to (approximately) 1923. As seen in the green rows of figure 1, SBLR was rerun twice, once by excluding all articles from the years 1920 to 1922 and then once by excluding all articles from 1920 to 1923. By no longer

requiring the model to predict for the specified years, the accuracy was improved to 79.7% for the first modification and 80.9% for the second.

We can justify the exclusion of the years by noting the lengthy duration of the Bolshevik Revolution. Since the revolution did not occur solely in a single year, and given that our experiment's purpose is to predict pre- and post- revolution years, the modification of SBLR to exclude the years 1920 to 1923 is justified.

The stepwise feature of SBLR is seen in figure 23, illustrating how the addition of variables affects the predictive accuracy. The model adds more significant variables with each successive step to improve the accuracy. It will continue to do so until no additional variable will significantly improve the predictive accuracy. This particular model was run with the exclusion of the years 1920 to 1922 with the years less than 1920 being pre and the years greater than 1922 being post. The model included 15 variables in 15 different steps where the other models (figure 22) took different numbers of steps to reach their highest accuracy.

Table 6. Stepwise Model Variable Addition and Accuracy Increase

Classification Table

Observed		Predicted			Variable Entered
		out1920_1922		Percentage Correct	
		<=1919	>=1923		
Step 1	out1920_1922 <=1919	581	30	95.1	topic14
	>=1923	279	146	34.4	
	Overall Percentage			70.2	
Step 2	out1920_1922 <=1919	572	39	93.6	topic05
	>=1923	236	189	44.5	
	Overall Percentage			73.5	
Step 3	out1920_1922 <=1919	566	45	92.6	topic08
	>=1923	214	211	49.6	
	Overall Percentage			75.0	
Step 4	out1920_1922 <=1919	573	38	93.8	topic20
	>=1923	205	220	51.8	
	Overall Percentage			76.5	
Step 5	out1920_1922 <=1919	571	40	93.5	topic10
	>=1923	198	227	53.4	
	Overall Percentage			77.0	
Step 6	out1920_1922 <=1919	569	42	93.1	topic39
	>=1923	191	234	55.1	
	Overall Percentage			77.5	
Step 7	out1920_1922 <=1919	565	46	92.5	topic37
	>=1923	189	236	55.5	
	Overall Percentage			77.3	
Step 8	out1920_1922 <=1919	558	53	91.3	FAC3_1
	>=1923	175	250	58.8	
	Overall Percentage			78.0	
Step 9	out1920_1922 <=1919	556	55	91.0	topic26
	>=1923	172	253	59.5	
	Overall Percentage			78.1	
Step 10	out1920_1922 <=1919	551	60	90.2	FAC19_1
	>=1923	160	265	62.4	
	Overall Percentage			78.8	
Step 11	out1920_1922 <=1919	551	60	90.2	topic11
	>=1923	161	264	62.1	
	Overall Percentage			78.7	
Step 12	out1920_1922 <=1919	549	62	89.9	topic36
	>=1923	159	266	62.6	
	Overall Percentage			78.7	
Step 13	out1920_1922 <=1919	544	67	89.0	topic02
	>=1923	157	268	63.1	
	Overall Percentage			78.4	
Step 14	out1920_1922 <=1919	544	67	89.0	FAC16_1
	>=1923	149	276	64.9	

	Overall Percentage				79.2	
Step	out1920_1922	<=1919	544	67	89.0	topic07
15		>=1923	143	282	66.4	
	Overall Percentage				79.7	

At step 15, the model is as follows:

Table 7. SBLR Excluding 1920-1922

	B	S.E.	Wald	df	Sig.	Exp(B)	
Step	topic02	-	3.731	8.124	1	.004	.000
15 ^o		10.633					
	topic05	10.161	1.668	37.106	1	.000	25875.707
	topic07	4.105	1.713	5.742	1	.017	60.659
	topic08	11.595	2.738	17.940	1	.000	108567.541
	topic10	-7.736	1.628	22.575	1	.000	.000
	topic11	15.532	8.980	2.992	1	.084	5565008.190
	topic14	25.895	3.505	54.577	1	.000	176228199102.066
	topic20	-8.978	1.649	29.650	1	.000	.000
	topic26	-5.890	1.465	16.156	1	.000	.003
	topic36	6.149	2.061	8.899	1	.003	468.268
	topic37	9.563	1.697	31.771	1	.000	14223.014
	topic39	18.628	6.258	8.860	1	.003	123098179.991
	FAC3_1	.621	.112	30.748	1	.000	1.860
	FAC16_1	-.202	.085	5.704	1	.017	.817
	FAC19_1	-.307	.097	9.971	1	.002	.736
	Constant	-.515	.184	7.807	1	.005	.597

The coefficients are given under the column B and the p-values are given under the column Sig. Topic 11 is of specific interest since its p-value is by far the largest, making it less significant than the other coefficients. However, under a 90% confidence level ($\alpha=0.1$), this coefficient is still significant resulting in a model that can be considered reliable. In terms of the actual coefficients, topics 8, 14, and 39 are most influential in determining the pre- post-binomial result within the model given their large magnitudes and extremely low p-values.

To interpret the results however, it is best to refer to Figure 2. By focusing on the order of inclusion into the model of the variables, we have an ordered list of the variables from most important to least. As such, the first variable, topic 14 is the most crucial. This topic includes

such words as: soviet, revolution, war, and communism. From a humanities perspective, the position of inclusion of this topic supports the validity of our regression model and our topic modeling since this topic is highly correlated with the Bolshevik Revolution.

Alongside this experiment, the effect of running the regression with “only topics” and then “only factors” was also tested, the results of which can be seen in Appendix J: Topic Modeling Experiment 1 Tables. “Only topics” regression was able to achieve a predictive accuracy of 79.4% in 17 steps while still choosing topic 14 first. However, many of the p-values were generally higher such as topic 11 which would not be significant under a 95% confidence level. “Only factors” produced a predictive accuracy of 76.4% in 17 steps. The p-values were all significant under the 95% confidence level. Overall, we see that “only factors” provides more significant coefficients while obscuring the specific topics that determine the article differentiation, a reasonable tradeoff.

The high accuracies of our models suggest that we can reject our initial null hypothesis (no difference pre- and post- sea change year). However, this is not to say that these results are without limitations. The predictive categorization of our explained variable into two separate, continuous time periods might be affected by influences other than the Bolshevik revolution. The change in language and content in US commentary could be influenced by a general change in language and biasing in our dataset. Before 1920, the intrinsic language independent of the topic ideas discussed may be different than the language after 1922. The natural change in popular language would contribute to the predictive differentiation of the categorical bins. Additionally, the concentration of articles about particular Russian authors changes over time, which may also contribute to the predictive differentiation.

4.2.2 Topic Modeling Experiment 2

We also consider SLBR's ability to predict a number of annotations questions. We include both SBLR (categorical) modeling and linear regression (continuous) modeling for the sake of testing the merit of SBLR.

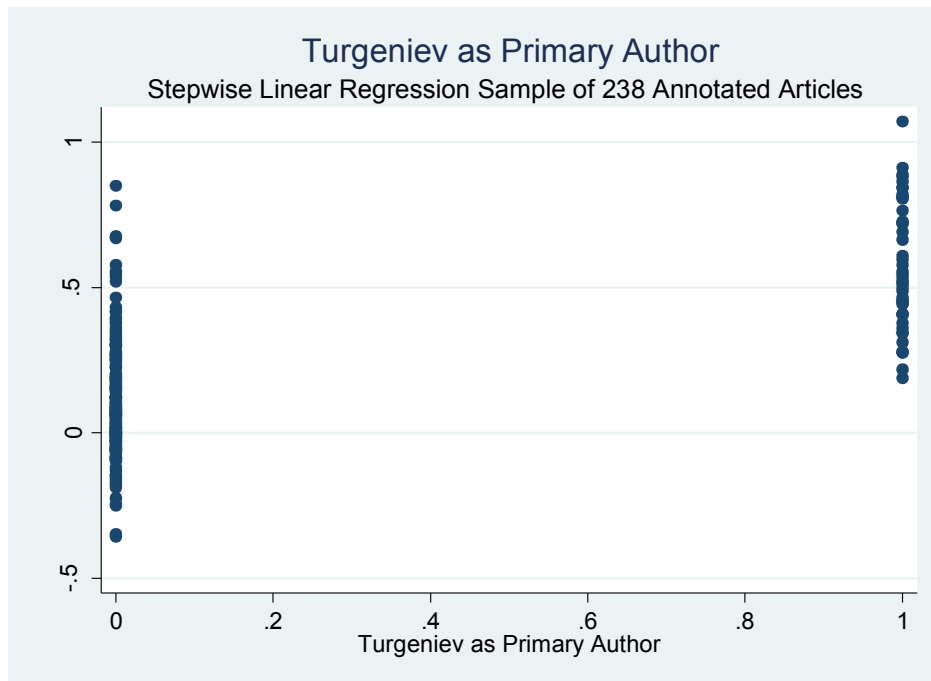
Out of the questions tested by WEKA, we focused on those regarding the identification of the primary author discussed, opinion of article writer towards the primary author, determination of radical politics as an issue, and prevalence of style. WEKA's accuracy for these questions was variable; some had very good accuracy and others were poor. Testing against these questions therefore gives a range of comparison for the assessment of topic modeling's effectiveness versus WEKA.

4.2.2.1 Authors

Within our annotated articles, 46 articles were identified as primarily about the Russian author Turgenev. Figure 22 shows predicted article values in a stepwise linear regression (from Equation (1)). Then, each article is assigned an outcome value predicting whether Turgenev is a primary author in that article. Since articles are binomial – either they are about Turgenev or not - interpretation of the linear regression outcome is fairly simple. The tool predicted that articles with a model outcome greater than 0.5 to be more likely about Turgenev than not. Our baseline accuracy rate of 79.8% is modeling off the ZeroR classifier; with the linear regression this accuracy rate rises to 87.2%.

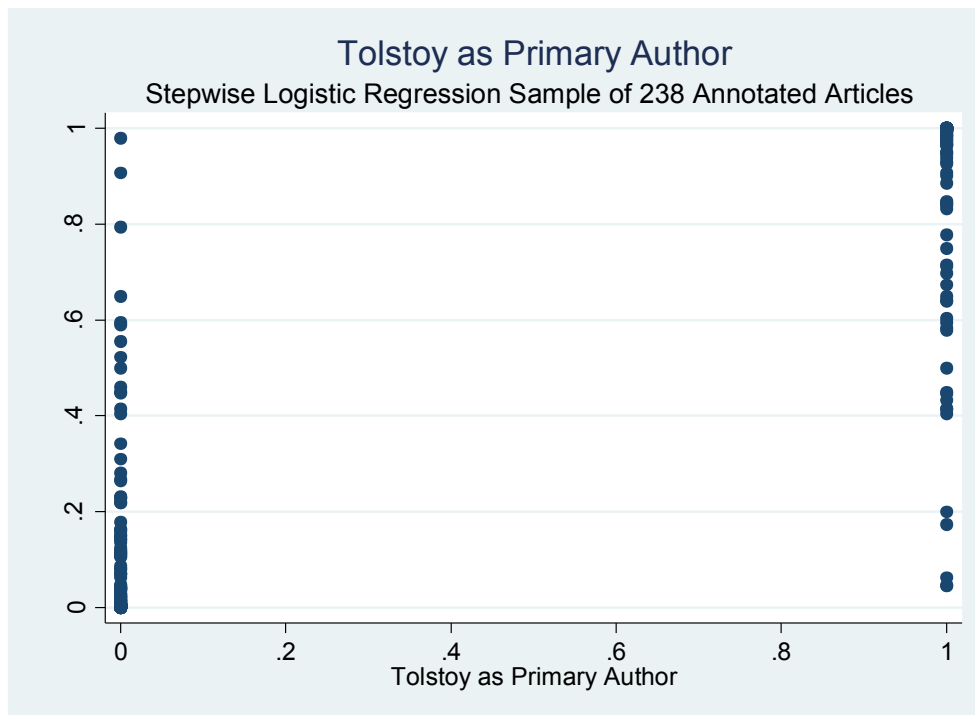
$$\text{Turgenev as Primary Author} = \alpha + \beta_0 \text{Factor}_i + \beta_1 (\text{Factor}_i)^2 + \epsilon_i \quad (1)$$

Figure 22. Predicted Article Values, Turgenev as Primary Author



With a SBLR model, the accuracy rate rises to 91.2%.

Figure 23. Predicted Article Values, Tolstoy as Primary Author



We also consider whether Tolstoy is a primary author with the same methodology in figure 23. 93 of the 228 annotated articles were about Tolstoy, which results in a baseline accuracy rate of 59.2%. Shown are both the linear model (figure 24) and the SBLR model (figure 25). The linear regression model improves accuracy to 90.3%, while the SBLR model improves accuracy to 91.2%.

Figure 24. Linear Model, Tolstoy as Primary Author

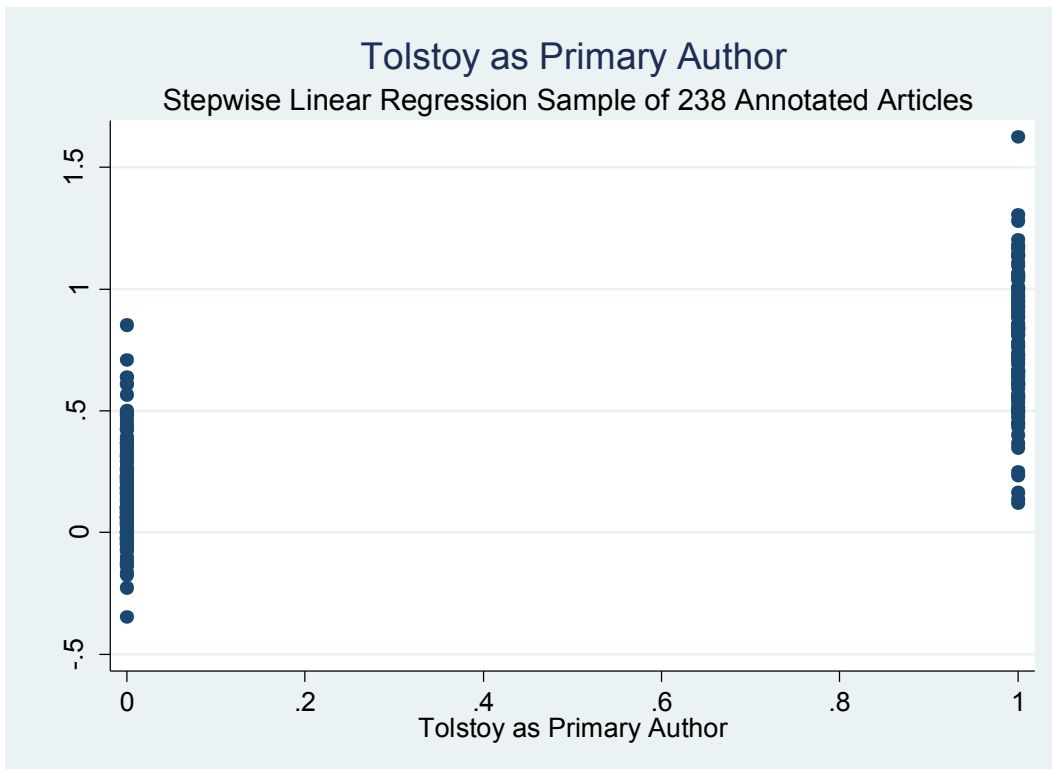
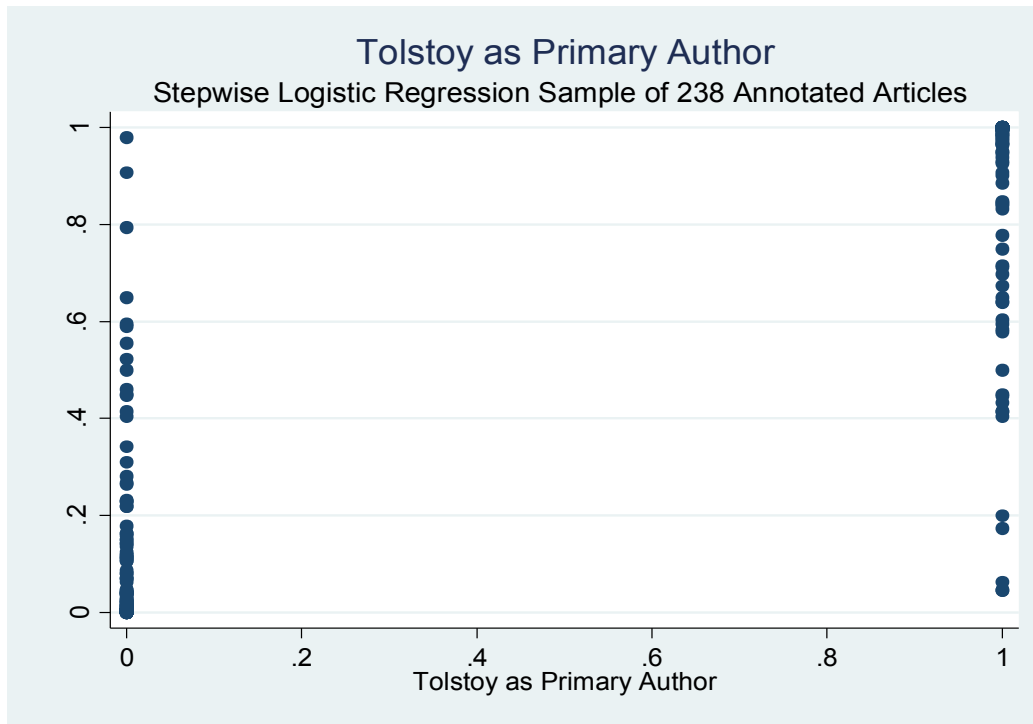


Figure 25. SBLR Model, Tolstoy as Primary Author

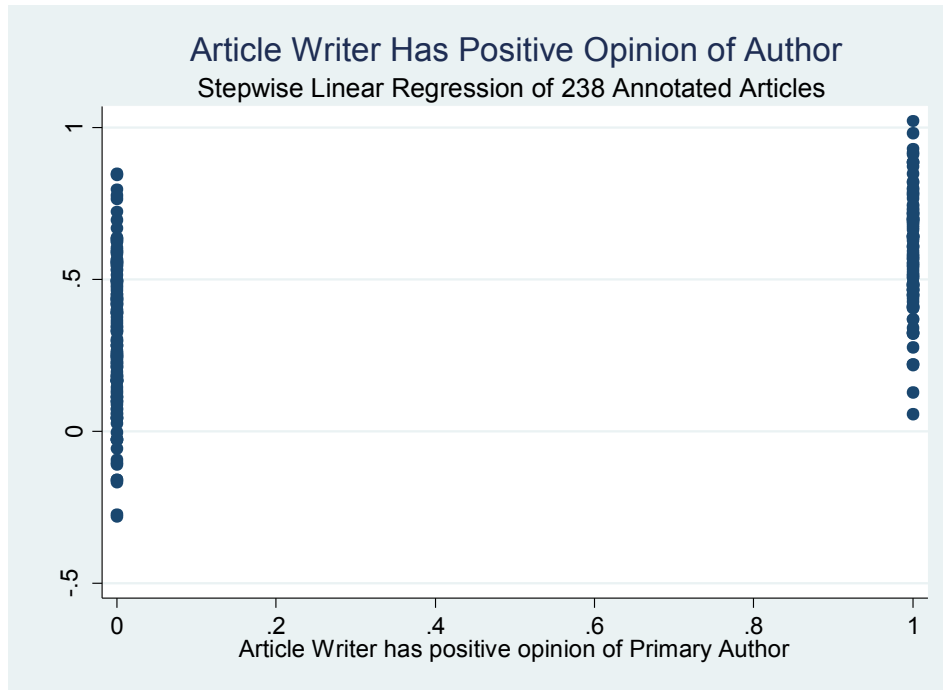


In many ways, high rates of improved accuracy over the baseline rate is somewhat expected because topic modeling appears to be able to differentiate between authors by dividing them into separate topics. This difference will appear in factors as well and in turn impact both linear and SLBR models predicting primary authors of articles.

4.2.2.2 Sentiment Analysis

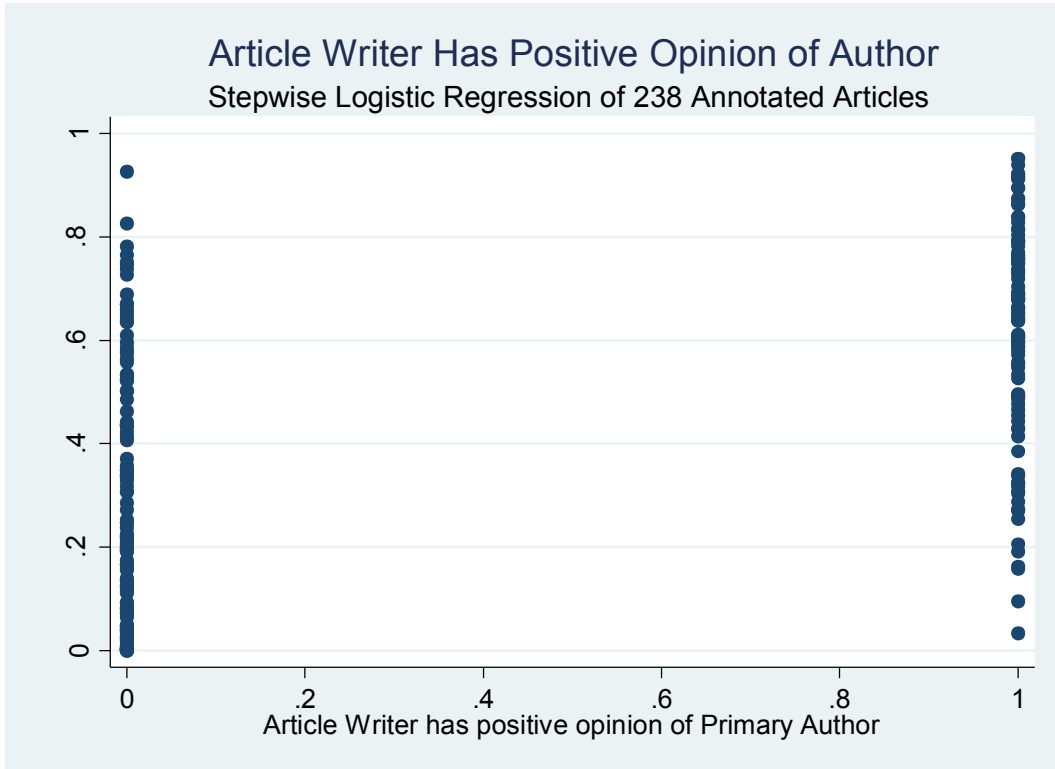
More challenging is accurately predicting article sentiment – it was a particularly challenging question for the annotation portion of the methodology as some articles were not explicitly positive or negative. We consider articles with positive sentiment – 97 of them – against all other articles. The baseline accuracy rate for positive sentiment is 57.4%. The linear regression model is similar to Equation (1) except with a dummy variable for positive sentiment substituting for primary author. Figure 26 contains outcome values for a linear regression model, which has an accuracy rate of 72.3%

Figure 26. Linear Model, Sentiment Analysis



Running the SLBR model for positive opinion, seen in figure 27, increases the accuracy rate to 73.2%. This is a significant improvement over WEKA's ability to predict positive sentiment.

Figure 27. SBLR Model, Sentiment Analysis



4.2.2.3 Radical Politics as an Issue

We also consider predicting whether radical politics is an issue. The ZeroR baseline accuracy for classifying an article as discussing radical politics is 65.6%. The linear regression model adjusts as it did previously. The linear model has an accuracy of 74.1%, depicted by figure 28.

Figure 28. Linear Model, Radical Politics as Issue

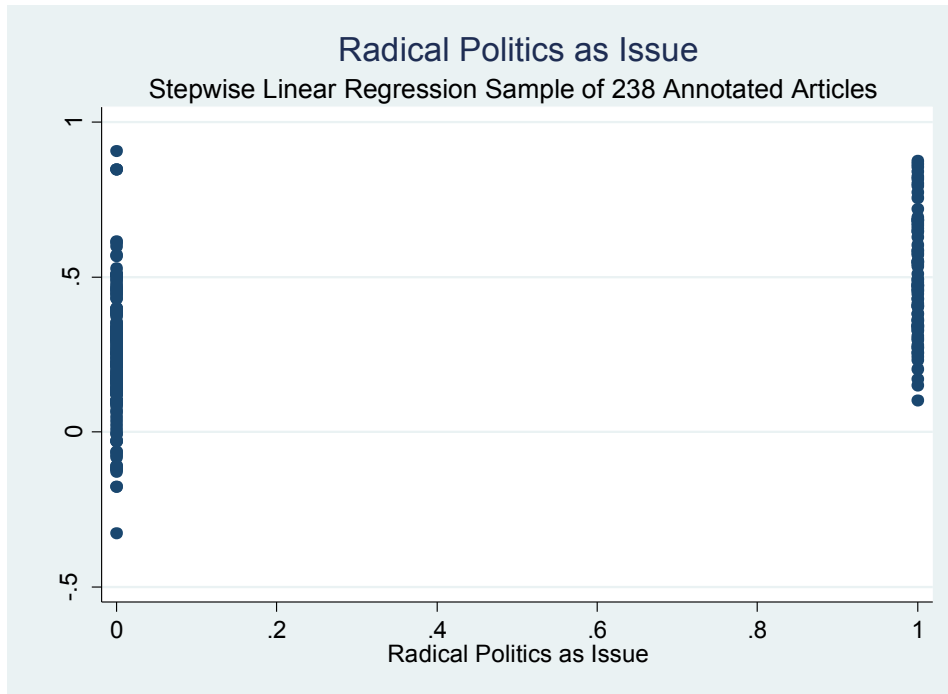
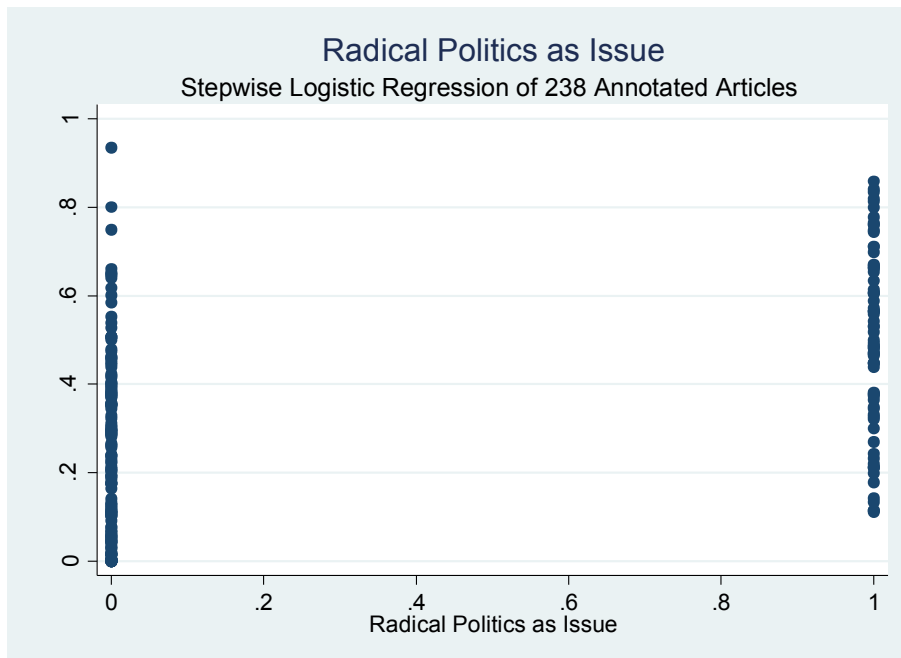


Figure 29 shows the SBLR model improved accuracy to 75.8%, which is consistent with the previous annotation experiments. Linear models are not suited to categorical regression mode.

Figure 29. SBLR Model, Radical Politics as Issue



Once again, the SBLR model's accuracy is an improvement over the WEKA accuracy. The presence of explicit radical politics topics likely aided the regression model. Table 8 shows the SBLR model coefficients output for radical politics.

Table 8. SBLR Model Coefficients Output, Radical Politics

	Logit Stepwise
N	228
Log Likelihood	-112.1
Factor 1	-0.615*
	(0.252)
Factor 2	0.856*
	(0.372)
(Factor 3) ²	-0.278**
	(0.095)
(Factor 7) ²	-0.988**
	(0.371)
(Factor 8) ²	-0.344**
	(0.131)
(Factor 9) ²	0.194**
	(0.063)
(Factor 6) ²	-0.443*
	(0.215)
Factor 9	0.674**
	(0.218)
Constant	0.327
	(0.256)

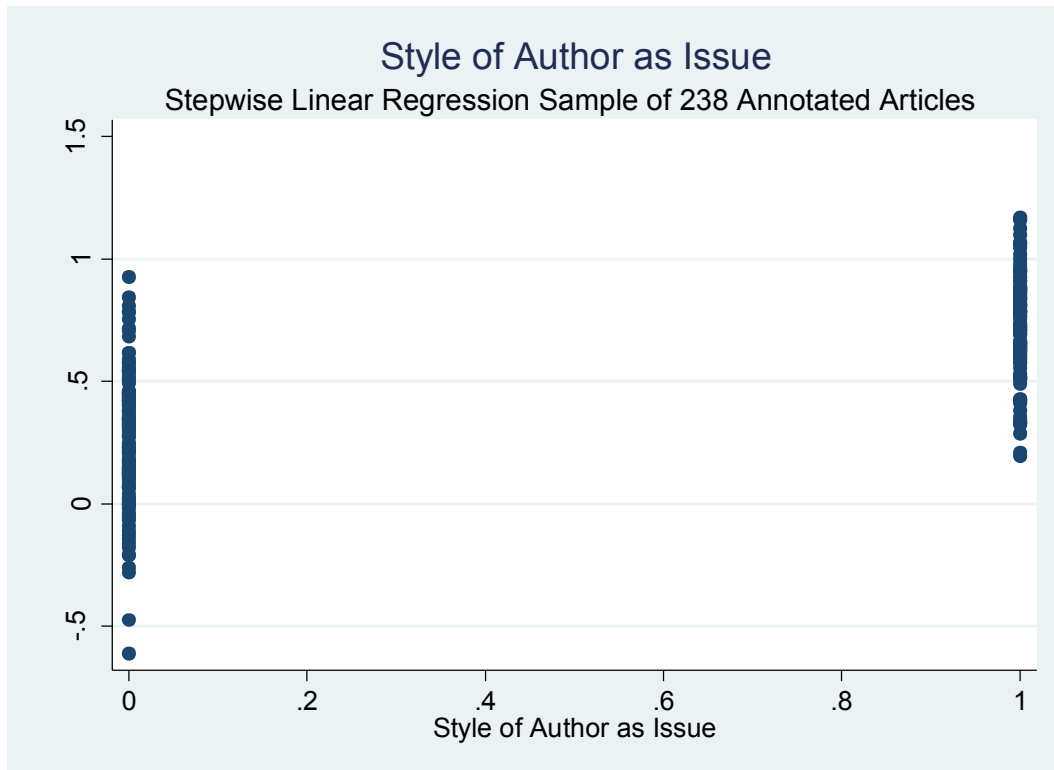
* p<0.05, **p<0.01, *** p<0.001

All coefficients except for the constant's coefficient are significant at the 95% confidence level due to the stepwise nature. It is also worth noting that unlike SPSS, STATA will automatically include squared values of variables in the regression. However, the order of the coefficients in STATA is not representative of their importance in increasing the predictive accuracy.

4.2.2.4 Style of Author as an Issue

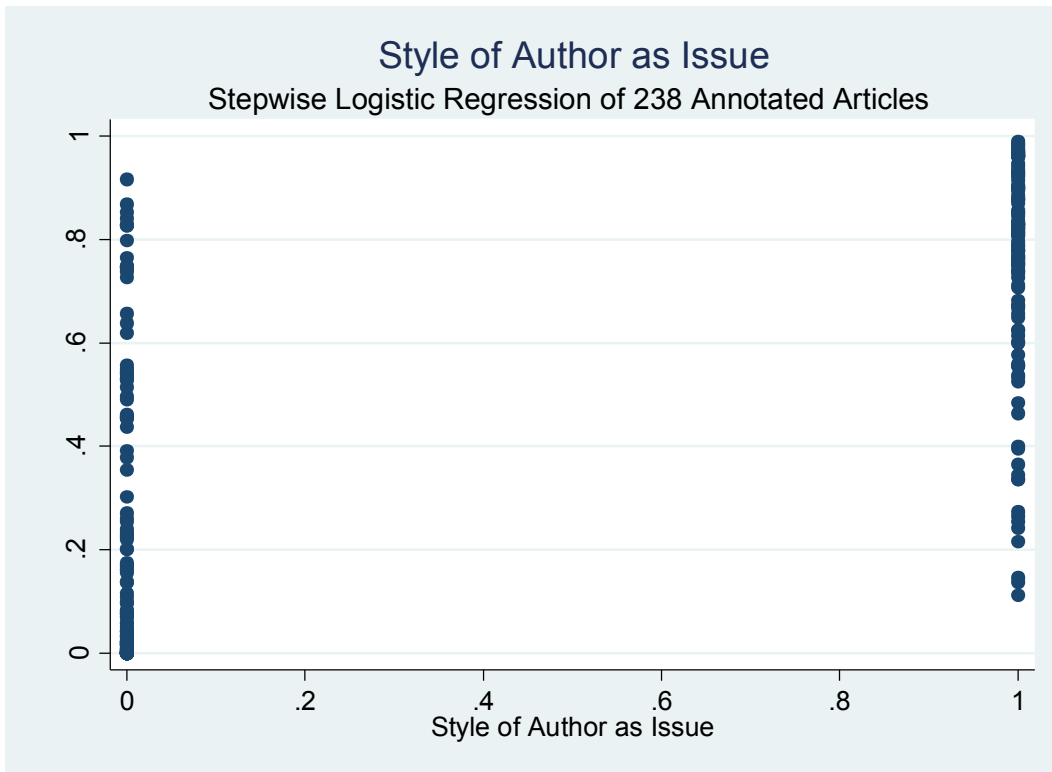
WEKA was less successful in predicting whether author style was an issue in annotated articles. The baseline accuracy for this question is 51.7%. Using a linear regression model, we improved accuracy to 83.3%

Figure 30. Linear Model, Style as Issue



This accuracy is a very large jump from the baseline. The logistic model's accuracy is 81.1%. As explained before, the logistic model should be the correct regression model for the demands of this experiment. To have this reversal in accuracy between the linear and logistic models suggests possible error.

Figure 31. SBLR Model, Style as Issue

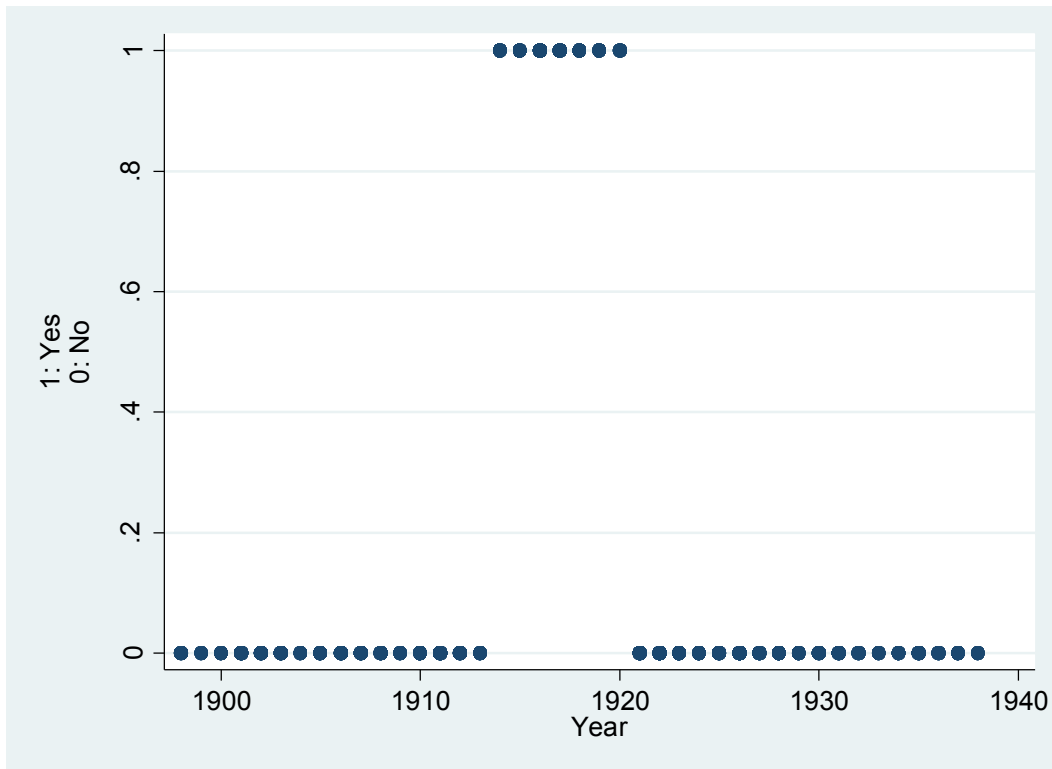


Overall, the Style annotation question may have been easier for the model to predict than WEKA because articles that deal with author style may be more likely to contain language dealing with putting the Russian author in a broader literary context. This could play a role in explaining the large jump from baseline accuracy to the model accuracies because topic modeling produced broader topics encompassing style whereas WEKA was limited to specific keywords. This intuition may explain the relative success of topic modeling versus WEKA across a number of annotation questions.

4.2.3 Topic Modeling Experiment 3

Given Experiment 1’s limitations, it is of interest to consider whether we could apply the concept of differentiating between time periods without suffering from a potential concern regarding long term changes in language. We consider whether we can apply our database to accurately predicting which articles occur during a period of crisis for Russia. As shown in appendix A, Russia’s involvement in WWI, and the subsequent Russian Revolution, intersects our database timeline. Figure 32 shows years in which Russia was in crisis during and after WWI – these dates include the years 1914 through 1920. 179 articles, or approximately 17 percent of the database, were written during these years. Predicting whether an article occurs during a crisis period in Russia’s history allows us to apply differentiation between time periods without the possibility that language in general changed slowly over time.

Figure 32. Russian Years of Crisis



We constructed a stepwise logistic probability model (SBLR) to determine which topic factors best predict that an article was written while Russia was involved in WWI and the Russian Revolution. Included in the logistic probability model are factors for each coefficient, as well as the squares of factors. Introducing squares of factors allows for the marginal effect of a factor to change.

Table 9 contains two stepwise logistic models, one that includes a term for each factor explanatory variable, and another that includes *factor*² explanatory variables as well. For the sake of conciseness, only the explanatory variables that are significant for either model are included. Column (2), which allows for changes in marginal effect of factors, has a higher log likelihood value, which suggests the model that includes *factor*² explanatory variables is better able to explain variance in the outcome variable (Russia being involved in WWI and its subsequent civil disorder).

Table 9. Crisis Year Logistic Models

	(1)	(2)
Factors	Y	Y
Factors2	X	Y
Log Likelihood	-422	-407.5
N	1013	1013
factor10	-0.452***	-0.707***
	(0.078)	(0.129)
factor9	0.344***	0.446***
	(0.091)	(0.104)
factor3	0.473***	0.544***
	(0.130)	(0.119)
factor4	0.308***	0.331**
	(0.090)	(0.102)
factor5	-0.273**	
	(0.094)	
factor15	-0.209**	-0.710***
	(0.065)	(0.175)
sqr_factor15		-0.213**
		(0.068)
sqr_factor10		-0.084*
		(0.035)
sqr_factor7		-0.105*
		(0.043)
sqr_factor3		0.115*
		(0.050)
_cons	-1.759***	-1.628***
	(0.099)	(0.117)

* p<0.05, ** p<0.01, *** p<0.001

We predicted an outcome of the column (2) SLBR for all 1013 articles. A higher value (closer to 1) corresponds with the article being more likely to occur during the designated crisis time period in Russian history. Articles that did take place during the years between 1914 and 1920 have significantly greater outcomes than articles that did not take place during the years 1914 to 1920. Interpreting the output of the SLBR model is relatively straightforward. Articles

with an SLBR outcome greater than 0.5 suggest that it is more likely to have occurred during the years of interest.

Table 10 shows equation 2 predicted outcomes for articles versus actual article metadata. The ZeroR base accuracy rate is 82.2% - the SLBR model improves accuracy only slightly to 83.3%.

Table 10. Predicted Outcomes for Experiment 3

		Actual		
		Russian Crisis Era (1914-1920)		
		0	1	Total
Russian Crisis Era (1914-1920) Predicted [SLBR]	0	822	157	979
	1	12	22	34
Total		834	179	1013

We can also think about predicted yearly averaged outcomes of the SLBR. In figure #, we group predicted article outcomes from the SLBR model by the year in which they were written. Graph 3 shows mean predicted article values by year and the 95 percent confidence interval for each year. As previously discussed the model is able to significantly differentiate between articles written during the time period of interest from those that are not. Figure 33 shows the percent of articles in each year that the model considered to have been written between 1914 and 1920. Interpreting the y-axis value is simple – a value closer to 1 means a greater percent of articles in the given year have a model outcome value above 0.5, and are therefore predicted to fall within the years of interest we predict in equation 2.

Figure 33. Predicted Article Outcomes by Year

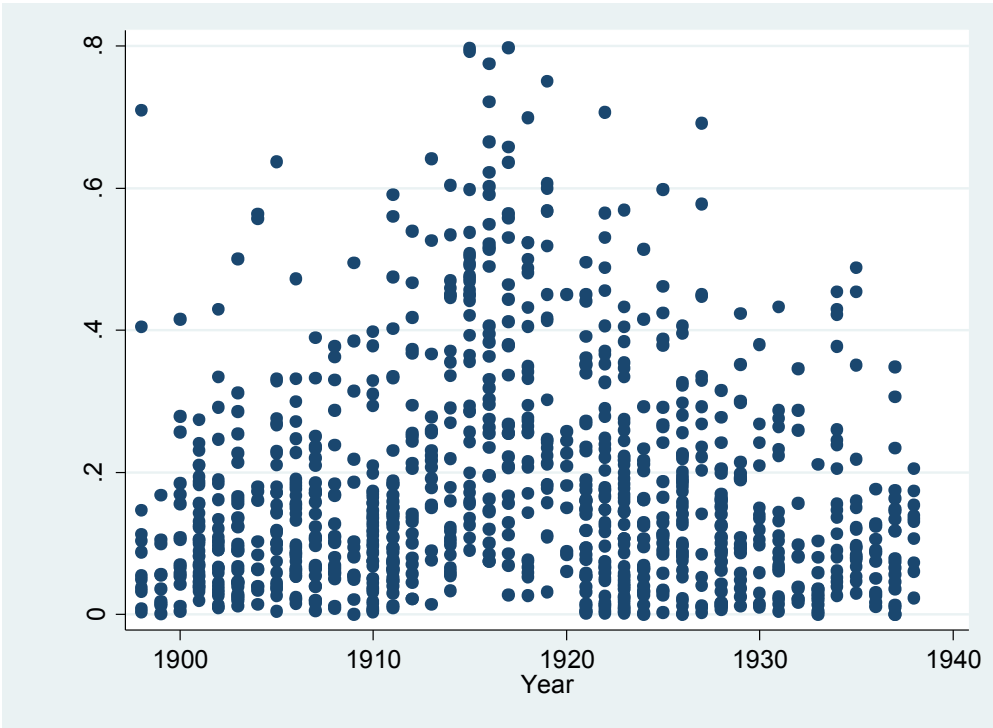


Figure 34. Mean Predicted Article Values by Year

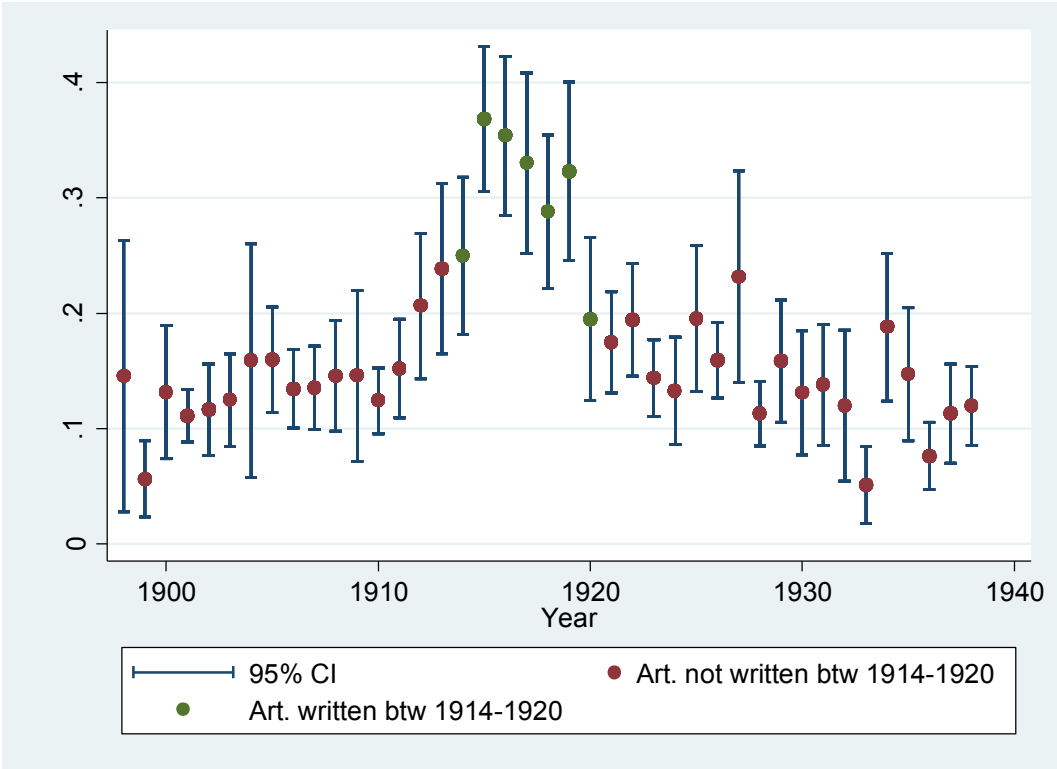
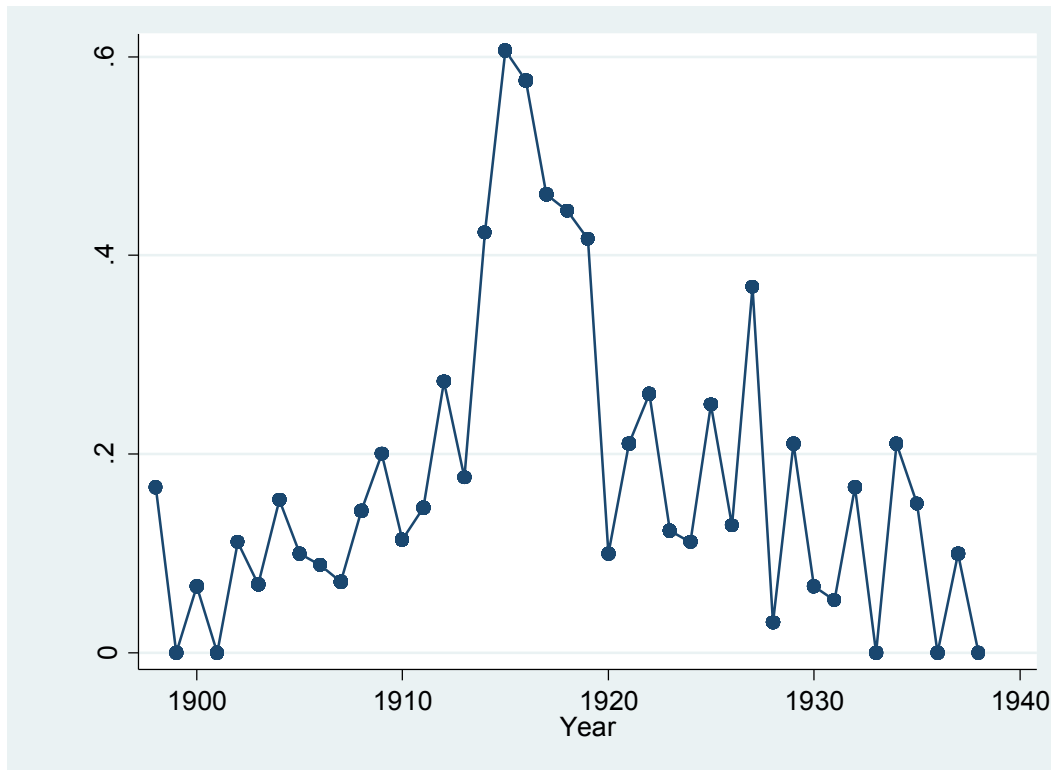


Figure 35: Percent of Articles Estimated to be Written During Crisis



Examining the mean predicted article values for the years 1912 and 1913, we see that visually these years appear to have different predicted means than in years prior to 1911. Individually comparing each year's mean predicted outcome value with all other years reveals significant differences between articles written during 1914-1919 and most other years. The year 1920 is not individually significant from years after 1920 according to the model.

When we grouped years together, as seen in Table 11, analyzing the difference between mean predicted values from the Equation 2 regression has similar results. Articles written between 1914 and 1920 have different mean predicted outcome values from almost all other year-groups (with an alpha of 0.05). Additionally, the two years leading to the crisis years of 1914-1920 are also significantly different from most non-crisis years; this is illustrated in Graph

3. This suggests that the model considers 1912 and 1913 to have some characteristics similar to 1914-1920.

Table 11. Percent of Articles Estimated to be Written During Crisis, Grouped Years

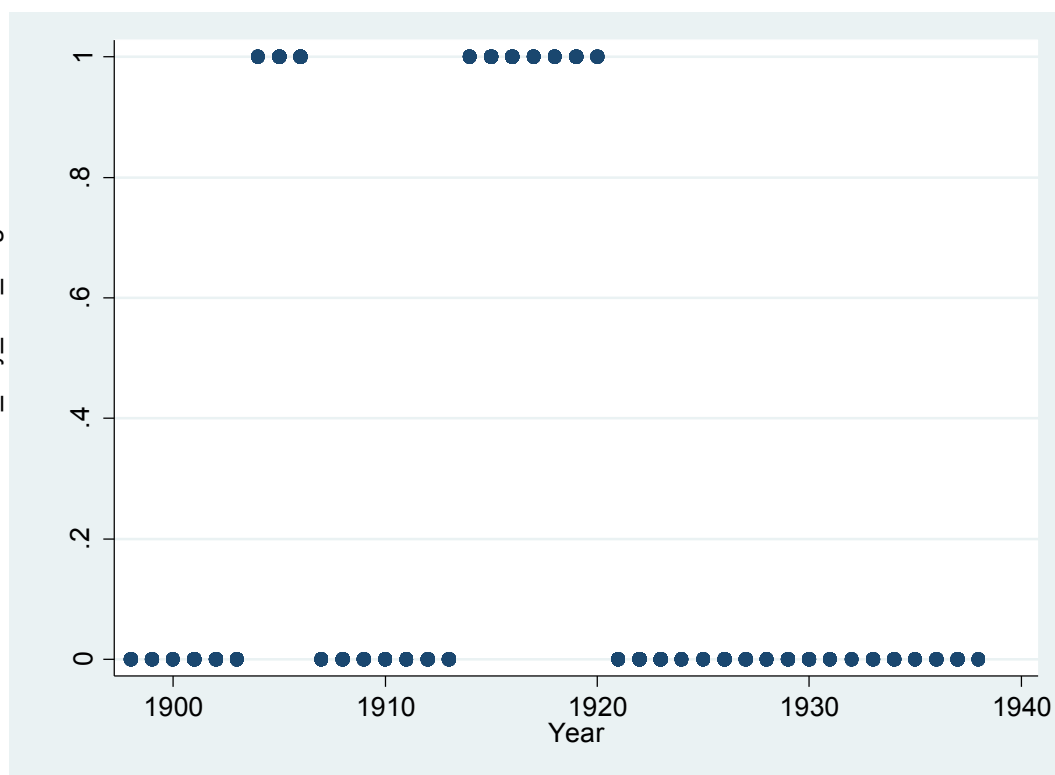
Years included in Group	Obs	% of articles predicted to be written between 1914-1920
1898, 1899, 1900	37	2.7%
1901, 1902	62	0%
1903, 1904	42	0%
1905	30	0
1906, 1907	62	0
1908, 1909	36	0
1910, 1911	85	1.1%
1912, 1913	39	5.1%
1914, 1915*	59	10.1%
1916*	33	21.2%
1917, 1918*	53	7.5%
1919, 1920*	34	14.7%
1921, 1922	84	4.7%
1923, 1924	84	1.1%
1925, 1926	63	1.5%
1927, 1928	52	1.9%
1929, 1930, 1931	53	1.8%
1932, 1933, 1934	43	0%
1935, 1936, 1937, 1938	62	0%

There are a few arguments that can be made as to why 1912 and 1913 are considered significantly more likely than pre-1912 years to have been written during the years 1914-1920. As previously discussed, the years 1914-1920 come at a significant time in Russia's history, including involvement in WWI and the Russian Revolution. The years 1912 and 1913 may be more significantly likely to occur during 1914-1920 according to our model because themes dominant during Russia's years of crisis were being discussed immediately prior to this time period – during the years 1912 and 1913. More broadly, one might argue that the significance of

1912 and 1913 comes from simple proximity in time – that similar events, authors, and books may have been discussed in 1912 and 1913, and therefore it is expected that a model which attempts to predict articles written between 1914 and 1920 will have some difficulty with ‘borderline’ years.

To address this potential issue, we consider predicting articles written during Russia’s involvement in any war or domestic conflict. Articles written between 1904-1906 correspond to armed revolts in Russia, as do articles written between 1914-1920 as discussed above. Figure 36 shows years at which Russia was involved in War or domestic conflict.

Figure 36. Russia at War or in Domestic Conflict



The new SLBR uses factor coefficients and squares of factors as explanatory variables to predict any Russian foreign or domestic conflict. Earlier, we were only attempting to predict if articles were written during a continuous group of years (1914-1920) – here we attempt to predict

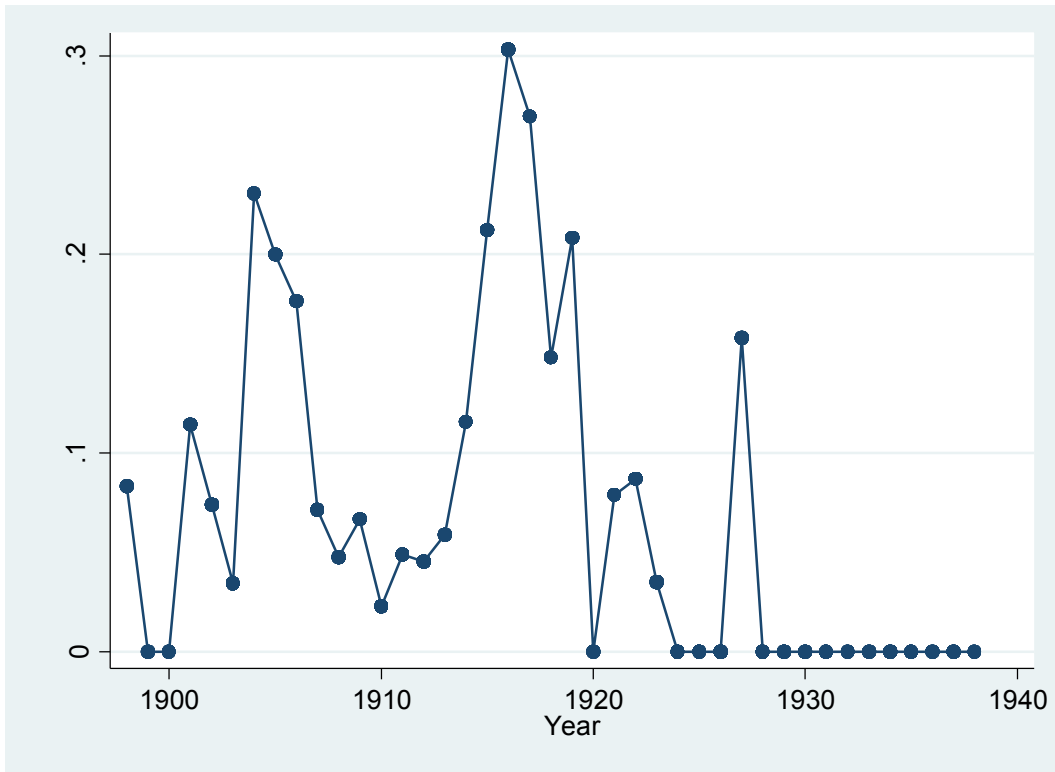
articles written during any Russian conflict, without the guarantee of continuity in predictive years. Graph 6 provides predicted values for articles according to the SLBR model graphed per year.

Figure 37. Predicted Values, SBLR Crisis Model



Figure 38 shows mean predicted article values by year, which shows peaks in predictive power during years of Russian domestic or foreign conflict. A value closer to 1 means a greater percent of articles in the given year have a model outcome value above 0.5, and are therefore predicted to fall within the years of interest we predict in equation (3). Many years have a mean predicted value of 0 because none of the articles within that year fall above the critical 0.5 cutoff which signals that the model believes an article was written during a crisis period.

Figure 38. Predicted Article Values by Year, Crisis Model



Comparing each year's mean predicted outcome value with all other years reveals significant differences between articles written during domestic or foreign conflict years versus most other years. Both the continuous groups of years between 1904-1906 and 1914-1920 are significantly different from other years.

Articles written between 1914 and 1920 as well as those written between 1904 and 1906 have different mean predicted outcome values from other years (with an alpha of 0.05). They also have higher mean predicted outcome values than articles written between 1907 and 1913. Comparing mean predicted outcome values per year against all other years reveals that years with Russian domestic and foreign crisis have significantly higher predicted outcome values than years in which Russia is not in a domestic or foreign conflict. As previously discussed, it is possible that general language trends are shared during all years in which there was Russian

domestic or foreign conflict. It may also be the case that certain Russian authors were more popular during the first half of our dataset, which could in turn drive the predictive power of the model upwards.

Running these SLBR models with topics excluding keywords associated with the Soviet Union yields a similar ability to differentiate between years of interest and years of non-interest.

Chapter 5: Conclusions

Team POLITIC sought to investigate methods with which to integrate data mining techniques in the digital humanities. Our hypothesis was that if data mining software and statistical tools were applied to a dataset of a corpus of periodicals, then new useful information could be uncovered for humanities research. The corresponding null hypothesis is that data mining techniques will only generate noise, due to the complexity of the unconventional data, and no new information can be gathered from the data. Our team believes that our experiments have produced meaningful, statistically significant results that add valuable information to Russian literature studies.

Team POLITIC researched the effectiveness of the two data mining techniques, text classification and Topic Modeling. While it is inappropriate to directly compare accuracy results from WEKA and regressions from Topic Modeling, it is promising that both methods saw increases in predictive accuracy. The fact that high predictability rates were achieved indicates that Topic Modeling produces data that is not dominated by noise and can be used to draw meaningful conclusions. WEKA was also used to some success in increasing predictive accuracy. Therefore, while more research must be completed to expand upon the conclusions acquired during this project, Team POLITIC feels confident that data mining techniques can be of use in humanities research.

In addition to the predictability rates, the models developed by the two data mining are enlightening and provide interesting insight. While WEKA's J48 classifier may not have been as successful as hoped, the words (attributes) chosen for the decision trees are often thought provoking. As stated earlier, it is curious that the J48 classifier would choose the words "surveillance" and "alarmed" as key attributes in the decision tree for the racial issues question.

This could spur the argument that article writers express caution and fear when discussing race during this time period. After analyzing the decision trees, many questions arise about the word choice, which could lead to future research directions.

Topic Modeling and the regression models, similarly, provide many opportunities for qualitative analysis that could prove fruitful. The words allocated to each topic are worth investigating because the Topic Modeling software generates the topics completely unsupervised. For our research, it is very reassuring that the generated topics appear to focus on themes that we would expect to see given the content of the dataset; topics about each of the major Russian authors, Russian revolutions, poverty, religion, literature and art. Team POLITIC spent significant time discussing the different topics and how the words in the topic are related. For example, Chekhov was frequently mentioned along with words related to theater and Dostoyevsky was frequently mentioned along with words related to Siberia and revolution.

The database created and data mining techniques investigated have demonstrated great potential for gathering useful data. Team POLITIC acknowledges that there are many improvements that can and should be made to our dataset and models, but believes the results presented in this paper warrant further consideration and inspire continued research.

5.1 Future Considerations

5.1.1 Annotations

As stated in Section 3.3.3 Selection of Annotation Articles, our sample was based mostly on convenience due to time constraints. Ultimately, this means that our sample of annotations may not be wholly representative of our final completed database. For better results, we suggest randomly selecting articles after database creation to input into the machine-learning program.

While the formation of our questions were created with the aim of simplification and to reduce error, there is still room for human error in our annotations. As stated above, we paired members of the annotation team with each other to compare their answers with the hopes that they would catch each other's mistakes. However, this method is not full proof as both members could make the same mistake. Also, even though we designed our questions such that the appearance of a key word would qualify for a "yes" answer, we did not have an all-inclusive list of key words so often times, annotation members would have to make their own decisions which would not necessarily been consistent with other annotation members. In the future we would like to address these inconsistencies and devise a better annotation questionnaire that our time restraints simply did not allow.

Furthermore, our questions were based solely on the issues we were interested in. We did not look at our dataset before considering our questions. As a result, many of our questions were very one-sided. For example, the question regarding the article author's gender had many more "male" answers than "female" answers simply because most of the critics at the time were male. Ideally, in order to optimize our results with machine learning, the sample dataset fed to WEKA would be balanced in the answers – having an almost even amount of "yes" answers and "no" answers. In the future we would suggest developing an overall understanding of the data first before devising questions so we can maximize the effectiveness of machine learning software.

Reading the articles with human interpretation would be inevitable although we addressed this with the explicit nature of our questions. However, there were some cases where our interpretations of an article would clash with the questions. For example, an article would mainly focus on the background of an author but a single mention of a religious word would automatically mean that we would answer "yes" to the "is religion an issue" question when this

would not necessarily be true. In the future we would like to change our annotation process to eliminate biases even more by having a few members who have a clear understanding of how to answer the annotation questionnaire create the entire dataset of annotated articles. This way the questions may not need to be as explicit and errors can be more readily caught.

Additionally, having a larger number of annotations would allow for WEKA to train on more annotations, and would possibly allow for greater increases in predictive accuracy. This would also allow for a larger sample size for SLBR, which is also promising.

5.1.2 Quantifying Foreign Policy

A major subset of experiments that we did not have the time and resources to run ourselves is analyzing how economic, militaristic, and political statistics from the time period correlate with our topic modeling dataset. We encourage future researchers to run these tests, as these could further show the versatility of our methodology, and could be more applicable to different research questions. The Chief of the Bureau of Statistics for the Department of Treasury publishes a series of books titled *The Foreign Commerce and Navigation of the United States for the Year Ending...* Each fiscal year ending June 30th has either one or two volumes detailing the exact goods imported from and exported to other countries.

The relevant economic statistics that we were interested in using included: total dollars traded with foreign countries, tonnage and number of vessels trading goods between the U.S. and foreign countries, and the amounts of specific commodities that were traded with foreign countries. The specific commodities were of particular interest to us, because the nature of these commodities could better convey the relationship of the U.S. with each country. For example, learning to which countries the U.S. exported gunpowder and gun blocks would provide us with concrete evidence of the U.S.' militaristic relationship with foreign countries. We would need to

speak to experts in foreign policy to better learn which traded goods are relevant to U.S. foreign policy.

Another dataset that is worth further investigation is immigration/emigration numbers. We hypothesized that increasing numbers of Russian immigrants to the U.S. would affect the U.S. perception of Russian authors, as determined by our annotations and topic modeling data.

Three other experiments we wanted to look at were the correlation of our topic modeling data with foreign aid statistics, with military statistics, and with political statistics. Foreign aid was hard for us to quantify, but seemed to us to be the most promising. One of our possible ideas was using food exported to Russia as the quantification of foreign aid. However, not all food traded is foreign aid. With more time, we would have researched which foodstuffs were most correlated with foreign aid, possibly by seeing which foodstuffs we most exported during times of famine. Military statistics were hard for us to come by during this period with Russia. Casualties and troop movements were the most obvious military statistics to examine. However, in our time period these are not relevant to U.S. interactions with Russia. For political statistics, one opportunity is use U.S. Communist party membership as an indicator for communist activity in the United States.

We were also unsure of how to best statistically analyze foreign policy data in relation to our topic modeling data. One of our ideas was to recreate the SBLR experiments using the foreign policy data to determine how to bin the articles. For example, we would bin each article based on whether trade with Russia was increasing (1) or decreasing (0) during the year it was written, and then run the same SBLR experiment. We were also interested in statistically analyzing whether regions of the U.S. with pro-Russian statistics (such as higher numbers of Russian immigrants, or higher proportion of citizens in the Communist party) would be the sites

of publishing more pro-Russian articles, as determined by the primary topics found in these articles.

5.2 Final Remarks

As the amount of data stored rapidly increases to unfathomable quantities, the ability to extract useful information from these databases is all the more paramount. Academic, commercial and governmental organizations will require data mining strategies. While Team POLITIC's research has focused on improving data mining strategies for the humanities and academia, we are excited at the prospect for future research directed for commercial and governmental organizations. The predictive power of the various models could gather very pertinent information in the present. As tensions heat up between the United States and Russia, it is not too fantastical to suggest that more refined data mining models might discover trends in United States discourse that indicate the potential for hostile engagements.

Appendices

Appendix A: Timeline of United States-Russian Relations

- 1763 - 1775: trade between Russia and British North America, against British law
 - American colonies began directly trading with Russia in 1763, in violation of Britain's Navigation Acts. This trade continued throughout the American Revolutionary War, which began in 1775.
- 1776 - 1781: though Russia, under Catherine the Great, remains officially neutral during US/Britain Revolutionary War, it leans toward the Americans, has interests in direct trade with the U.S. that presuppose the revolutionaries winning, and Russian systematically refuses a series of attempts by Britain to join an alliance against the US
 - March 1780 Russian ministry issues "Declaration of Armed Neutrality" (actually quite favorable to Americans in terms)
 - October 1780: Russia tries to mediate peace among European powers concerning US revolutionary war
- 1801: T Jefferson appoints Levett Harris first consulate- general to Russia (i.e., diplomatic relations between U.S. and Russia first established)
 - July 14, 1809: US first established diplomatic relations with Russia
- 1815: Congress of Vienna, Russia gains Kingdom of Poland (R. Divide)
- December 2, 1823 – Monroe Doctrine, warning to European countries against dabbling in affairs in the Western Hemisphere, but was in reality mostly directed against Russia
 - After Napoleonic Wars of 1803-1815, Prussia, Austria, and Russia formed Holy Alliance to defend monarchism. Specifically tried to reestablish the House of Bourbon (French) rule over Spain's colonies, which were to become independent
 - In addition to making the original statement, US made a second statement directed at namely the Holy Alliance stating that it is opposed to interpositions that would create new colonies among the independent Spanish-American republics
- 1830-1: November Uprising of Poles against Russian rule (R. Divide)
- 1854-6: Crimean War (R. Divide)
- 1855-81: Reign of Alexander II (R. Divide)
- 1860: State Bank founded (R. Divide)
- 1860's: arguable beginnings of revolution in Russia: "The revolutionary movement became an intrinsic element of Russian history as early as the 1860s" (xxi)
- 1861-1876 – Great Reforms to reform Russia's social and economic structure in the wake of the stunning defeat in the Crimean War
- 1861-65: Alone among European Powers, Russia offers rhetorical supports for Union during U.S. Civil War (key concern: U.S. counterbalance to British Empire)
 - Note: Emancipation of Serfs in 1861 coincident with Emancipation Proclamation
- 1862: Turgenev, Fathers and Sons (R. Divide)
- 1863: Chernyshevsky, What Is to Be Done? (R. Divide)

- 1863: University statute reforms Russian higher education (R. Divide)
- 1863-4: January Uprising of Poles against Russian rule (R. Divide)
- 1864: Dostoevsky, Notes from the Underground, Demons (R. Divide)
- 1865: General Michael Cherniaev takes Tashkent (R. Divide)
- 1865-9: Tolstoy, War and Peace (R. Divide)
- 1866 – Russian-American trading company collapses
 - Had the monopoly on trade in Russian America, which extended to the 55th parallel, and included the Aleutian Islands
 - Created settlements in modern-day Alaska, Hawaii, and California.
 - Alaskan purchase in 1867 gave Alaska to U.S., sold commercial interests
- 1867: U.S. purchases Russian America-- (i.e., Alaska) --from Russians
 - key concern of Russians is that this territory would not fall into British hands
- 1869: N. Danilevsky, Russia and Europe (R. Divide)
- 1871: Vereshchagin, “The Apotheosis of War” (R. Divide)
- 1872: Special higher education courses for women set up in Moscow (R. Divide)
- 1875: Uniates in Russian Empire converted to Orthodoxy
- 1875-7: Tolstoy, Anna Karenina (R. Divide)
- 1877-8: Russo-Turkish War (R. Divide)
- 1878: Congress of Berlin (R. Divide)
- 1881 – Bombing of Tsar Alexander II by the People’s Will revolutionary organization
 - regulations on students and universities ensue
- 1881: Attacks on Jews (“pogroms”) in Southwest (Ukraine) provinces (R. Divide)
- 1881-94: Reign of Alexander III (R. Divide)
- 1883: State Peasant Land Bank established (R. Divide)
- 1884: Repin, “They Did Not Expect Him” (R. Divide)
- 1885: State Noble Land Bank established
- 1891-1892 – Russian Famine, Americans sent ships loaded with flour
 - Tolstoy blamed famine on Tsar and the church
 - Government received widespread blame and discredit
 - Future Tsar Nicholas II aided in relief efforts
- 1892: Levitan, “The Vladimirka” (R. Divide)
 - Vladimirka was a road leading from Siberia to Europe, often used to traffic crowds of prisoners sentenced to exile.
 - The penal function that the road served figures into works of Herzen, Nekrasov, and Dostoevsky (Crime and Punishment)
 - After Russian Revolution, Bolsheviks were keen to get rid of the notorious rep of this road so they changed the name to Shosse Entuziastov
- 1893 – Founded the Social Democratic Labor Party
 - united revolutionary organizations

- Based on Mark and Engels theories of the working class
- members arrested at first congress in 1898
- Later splintered into Bolshevik and Menshevik
- 1894
 - Franco-Russian military advance (R. Divide)
 - Triple Alliance left Russia vulnerable and France was politically isolated after its defeat in the French-Prussian war
 - Franco-Russian alliance finalized Jan 4, 1894
 - Designed to defend against the Triple Alliance
 - Crucial piece in contributing to WWI
 - July: Free Russia newspaper in America ceased publication due to low subscriptions
 - August 1, 1894 – April 17, 1895 – Sino-Japanese War
 - fought between Qing Dynasty China and Meiji Japan in Korea.
 - After Japan fought off Chinese influence in Korea, Russia, Germany, France forced Japan off the Korean peninsula (Port Arthur)
 - After Japan left, Russia installed a king of the Russian legation to rule in Korea and (1898) signed a 25-year lease to Liaodong Peninsula as well as Port Arthur and built a railroad from St. Petersburg to Port Arthur
 - In 1900, Boxer rebellion, Russia took hold of Manchuria and installed troops in the area in hopes of gaining more influence in the Far East
 - November 1: Tsar Alexander III dies and is replaced by son Nicholas II. Russophilia started to resurge in America
 - Mid-1905, Nicholas II accepts American mediation, end to Russo-Japanese war
- 1894-1917: Reign of Nicholas II (R. Divide)
- December 10, 1895 – Russo-Chinese bank created
 - Bank founded in St. Petersburg, representing Russia's interest in China
 - Notable Supreme Court Case: Russo-Chinese Bank v. National Bank of Commerce
 - Some issue about how the Russo-Chinese Bank was asking a Seattle bank for payment for shipping document, but jury ruled that Seattle bank already paid.
- June 1896 – Russo-Chinese bank set to construct Manchurian railroad
 - Named Chinese Eastern Railway, linking Manchuria with Vladivostok
 - After first Sino-Japanese gained right to build this and had a large army that occupied Northern Manchuria.
 - Russia pressed China for a “monopoly of rights” in Manchuria, to which China responded by an alliance with Japan and United States against Russia
- Jan. 1897: Russia adopts gold standard (R. Divide)
 - introduced gold standard as a means to attract foreign capital in order to sustain ambitious industrialization plans, and to earn respectability since most superpowers had adopted gold standard
- April 1898 – Start of Spanish-American War

- Point of contention with Russia: US acquired the Philippines after the Treaty of Paris
- Russia wanted to help Spain in the retention of the Philippines, in hopes of that they may serve as a Russian food base in the Pacific
- December 1898 – US Acquisition of Philippines
- February 1899 – Statement by Russian Committee Chairman expressing tension
- February 1899: outbreak of large scale unrest and strike at Russian universities
 - “these disorders set in motion a movement of protest against the autocracy that did not abate until the revolutionary upheaval of 1905--6” (4)
 - On the anniversary of the founding of St. Petersburg Univ. (Feb 8), students would party in Nevsky Prospekt.
 - 1895, Students/Janitors Brawl; 1897, 500 students march on Winter Palace for public dance; 1898, students do same thing and resisted police this time
 - 1899 Ministry of Education banned street parties and would be arrested
 - On Feb 8, police blocked bridge leading to city center, mounted police ambushed students and responded to student snowballs with whips
 - US sides with students
- 1899: First Hague Convention called by Nicholas II (R. Divide)
 - Proposed Aug 29 1898, Signed July 29.
 - Created Permanent Court of Arbitration, ratified by major powers including US
 - provides services of arbitration and resolution of disputes between states
 - Conventions with respect to conventions and customs of war
 - Conventions of maritime warfare and principles
 - Invoked brownie points between US and Russia
- 1900: U.S. and Russia allied during Boxer Rebellion (defeating Qing rebels); Russia had occupied Manchuria at this time
 - 8 nations joined to quell rebellion of Boxers
- 1901: Lev Tolstoy excommunicated from the Orthodox church (R. Divide)
 - Tolstoy began concentrating on Christian themes (The Death of Ivan Ilyich, What is to be Done), radical anarcho-pacifist Christian philosophy, led to excommunication
 - radical anarcho-pacifist - rejects use of violence for social change
- 1902 – Socialist Revolutionary Party founded (most radical: combines anarchism, syndicalism, terror; three main planks: anti-capitalism, terrorism, socialization of land))
 - key player in the first Russian revolution
 - garnered much support from peasants - division of land to peasants v. collectivization in state management
 - Believed that the laboring peasantry as well as the industrial proletariat will be driving force in revolution
- Russian Social -Democratic Labor Party: after abortive start in 1898, comes in existence at its second congress in Belgium and England in 1903 (no terror at this point: appeals prim to industrial working class, and also as a provisional measure bourgeoisie)

- Split into Mensheviks and Bolsheviks at the second congress on Nov 17
- 1903: Prohibition of printing Lithuanian in Latin letters dropped (R. Divide)
 - ban on all Lithuanian language publications printed in the Latin alphabet within the Russian Empire, which controlled Lithuania at the time
 - Tsarist hoped that this would decrease Polish influence in Lithuania and return Lithuania to its ancient roots, to Russia
- Spring 1903 – Hundreds of Jews killed at pogrom in Kishinev, sparked riots throughout US. President Roosevelt forwarded a petition from the American people to Russia
 - 49 Jews killed and hundreds wounded
 - Jews in US organized nationwide protests.
 - In addition to hundreds of protests and demonstrations, organized massive petition
- 1904: Ivan Pavlov receives Nobel Prize for Physiology-Medicine (R. Divide)
- 1904-05 February 8 1904 - May/June 05: Russo-Japanese War
 - a key thing to note about this war: it keeps the Russian army away, far on the other side of the continent, and devastated during the domestic upheavals of 1905; much domestic instability if consequent in this power vacuum. Also hugely assails the legitimacy of the Tsarist monarchy.
 - Feb 20, 1905 to March 10 1905: big defeat for Russians in Mukden
 - Last and decisive battle of the Russo-Japanese war
 - 340,000 Russian v. 280,000 Japanese
 - Pushed Russia out of southern Manchuria for good
 - Shocked powers of imperial Europe since Russia had more manpower, materials
 - Proof that Europeans not invincible, could be decisively outmatched in battle
 - Russian empire shifted policy toward Balkans, would eventually lead to WWI
 - May 14-27, 1905: greatest naval defeat in Russian history at hands of Japanese in the Strait of Tsushima
 - nearly entire fleet destroyed while Japanese only lost 3 torpedo boats
 - June 1905 US President Theodore Roosevelt mediates peace talks in New Hampshire
 - Both nations agreed to retreat out of Manchuria
 - Set Pacific balance of power for Russia, Japan, China, Korea, Europe and US
 - Japan becomes world power
 - US becomes world leader in diplomacy
 - Roosevelt receives Nobel Peace Prize in 1906
 - Sept 5, 1905: Treaty of Portsmouth: “Russia surrendered the southern half of Sakhalin and consented to Japan’s acquiring the Liaotung Peninsula with Port Arthur, as well as establishing hegemony over Korea, neither of which were Russian property. There was to be no indemnity. The price was small, considering Russia’s responsibility for the war and her military humiliation” (35).
 - Useful (implicates question of race, foreign policy, and Russia): “Russia’s defeat at the hands of the Japanese was to have grave consequences for the

whole of Europe by the lowering the esteem in which whites had been held by non-Western peoples; for it was the first time in modern history that an Asiatic nation defeated a great Western power. One observer noted in 1909 that the war had ‘radically reshaped’ the mood of the Orient: ‘There is no Asiatic country, from China to Persia, which has not felt the reaction of the Russo-Japanese war, and in which it has failed to wake new ambitions. These usually find expression in a desire to assert independence, to claim equality with the white races, and have had the general result of causing Western prestige to decline in the East’ (Thomas F. Millard, -America and the Far Eastern Question- New York, 1909, 1-2). The war marked the beginning of the process of colonial resistance and decolonization that would be completed half a century later” (35)

- 1905
 - October 17: Nicholas signs October Manifesto promises political reform (R. Divide)
 - Lenin, “Socialism and Religion” (R. Divide)
 - Periodical publications in Yiddish and Ukrainian allowed (R. Divide)
 - October: Constitutional Democratic Party (liberals, but with leftwing orientation)
- 1905 – Peace Talk conferences for Russo-Japanese War, Russia concedes Manchuria
 - Treaty of Portsmouth, signed in Maine
 - Arbitrated by Teddy Roosevelt, allowed Tsar’s refusal to pay compensation to Japan
- 1905: First Russian Revolution
 - domestic violence herein is arguably “the first phase of the Russian Revolution in the narrow sense of the word” (xxi)
 - “This First Revolution was also eventually crushed but at a price of major political concessions that fatally weakened the Russian monarchy” (4)
 - Struve on Jan 2, 1905: “In Russia, there is as yet no revolutionary people” (21)
 - changes with massacre of worker demonstrators in St. Petersburg on January 9
 - Jan 9: Bloody Sunday massacre: “spread the revolutionary fever to all strata of the population and made the Revolution truly a mass phenomenon” (21)
 - Bloody Sunday: (26) “among the masses, it damaged irreparably the image of the ‘good Tsar.’”
 - 1905 – Bloody Sunday massacre in St. Petersburg as a panicked and violent reaction to petitioners on the part of the tsarist police
 - January 22nd, peaceful demonstration to petition Tsar Nicholas II, led by Father Gapon, patriotic and religious march
 - Guards fired near Winter Palace, number killed and injured not known, but could be in the thousands
 - Brought bitterness to Tsarist regime; emotionally affected Leo Tolstoy
 - “In January 1905 over 400,000 workers laid down their tools: it was the greatest strike action in Russian history until that time” (26)
 - May 8, 1905: formal federation, instantiation, of the Union of Unions (30)

- August 6, 1905: “Bulygin Constitution” released for discussion
- October 10--17: crisis week w/massive strikes, culminating in October Manifesto on the 17th (big liberal concessions by Tsar) (44)
- 1905-06: agrarian revolt: peasant response to Manifesto: thru 1907 these are years of intense friction between reactionary and democratized forces. Agrarian revolt was less homicidal than peasant attempt to get landlords to abandon their property and sell their land at bargain prices
- Dec 6, 1905 Moscow Rising (socialist sponsored)
- Final comment: “The year 1905 marked the apogee of Russian liberalism—the triumph of its program, its strategy, its tactics. It was the Union of Liberation and its affiliates, the -zemstvo movement and the Union of Unions, that had compelled the monarchy to concede a constitutional and parliamentary regime. Although they would later claim credit, the socialists in general and the Bolsheviki in particular played in this campaign only an auxiliary role: their one independent effort, the Moscow uprising, ended in disaster.” (51)
- 1905 – Tsar’s Easter edict on religious toleration seen as first step towards redeeming Russia
- October 1905 – Tsar Nicholas II signs a reform manifesto allowing free speech, free Parliament, and freedom of conscience
- December 1905 – News of armed revolts in Moscow and St. Petersburg reach the US
- 1906: inaugurates Constitutional era (unstable): (159) “In some respects, perhaps the single most important prerogative of the new parliament was its members’ right to free speech and parliamentary immunity. From April 1906 until February 1917, the Duma provided a forum for unrestrained and often intemperate criticism of the regime. This probably contributed more to undermining the prestige of the Russian Government in the eyes of the population than all the revolutionary outrages, because it stripped the establishment of the aura of omniscience and omnipotence which it strove so hard to maintain.”
 - March 4: Laws issues guaranteeing rights of assembly and association
 - April 26: New Fundamental Laws (= Constitution) made public-- contradictory document simultaneously creating parliamentary Duma and licensing autocratic monarchical Czar
 - April 27: Duma opens (though First Duma is dissolved on July 8, not to be reopened until Feb 20 of following year; subsequent dissolution results in third opening on November 7, 1907)
- 1906-9: “Pig War” between Serbia and Austria (R. Divide)
- 1906-11: Peter Stolypin Prime Minister (R. Divide)
- 1906-17: Duma period (R. Divide)
- July 1906- -- April 1911: P.A. Stolypin is Russian Prime Minister
- 1906-1907 – Jewish Migration away from Russia
- 1907 – Russia signed the Russo-Japanese Treaty
- 1907-1912 – Pastor Boettcher, president of Russian Adventist union, converts thousands

- February 1910 – Russia formally rejected the US’s Chinechow-Aigun Railway plan
- 1910 – Russo-Japanese Treaty to work with Japan against the American railway plans
- 1911 – Russia doesn’t permit the passports of American Jews
 - 1911 – Russia doesn’t permit the passports of American Jews
- 1911: December: U.S. Senate, on recommendation of Pres. William Howard Taft, unanimously renounces US--Russian Treaty of 1832 (a commerce and navigation treaty) in objection to repeated refusal of Russian authorities to grant entry visas to American citizens of Jewish faith (178)
 - Significant instance of how Russian anti-Semitism poisons foreign relations with US
- 1912: Conclusive split occurs between the Bolsheviks and Mensheviks in Russian Social Democratic Party.
- July 1912 – Secret Russo-Japanese treaty regarding the splitting of Mongolia into East/West (Russia/Japan respectively) is signed.
- 1912 - 1916 – Russian government begins outlawing Christian missions, such as Adventist and Methodist missions.
- March 1913 – US President Wilson withdraws government support from the Manchuria consortium, ending major US involvement in China.
- July 1914: Russian Army mobilizes, and Germany declares war on Russia.
- July 28, 1914 - November 11, 1918 – World War I occurs.
 - World War I was a global war fought between two opposing alliances.
 - One alliance was the Allied (Entente) Powers, which originally consisted of the Triple Entente of the UK, France, and the Russian Empire.
 - The other alliance was the Central Powers, which originally consisted of Germany, and Austria-Hungary.
 - Both alliances expanded to include other powers throughout the war.
 - The US joined the Allied Powers in 1917.
 - End of WWI: German, Russian, Austro-Hungarian, and Ottoman Powers dissolved.
- March 15, 1917 – Tsar Nicholas II of Russia abdicates.
- February - November 8, 1917: Russian Revolution occurs.
 - February 1917: A mutiny of the Petrograd military garrison occurs.
 - February - March 1917: Revolutionary violence resumes. Tsarism collapses.
 - March 8 - 12, 1917: February Revolution occurs.
 - --March 9: US recognizes Provisional Government.
 - October 1917: Bolshevik coup d’état occurs.
 - November 7, 1917: Normal US diplomatic relations with Russia are interrupted.
 - November 7 - 8, 1917: October Revolution occurs.
 - December 6, 1917: Wilson orders all American representatives to have no communication with Bolsheviks. Although diplomatic relations are not formally severed, the US refuses to formally recognize or have formal relations with Russia/Soviet Union until 1933.

- 1917 - 1920: Bolshevik Revolution occurs.
 - 1917: First overthrow of the autocracy occurs.
 - 1919: Second revolution occurs due to struggles between rival political parties
- January 8, 1918: Woodrow Wilson's Fourteen Points Speech-- expresses concern about Russia.
- March 3, 1918: The Treaty of Brest--Litovsk was signed, resulting in Russia exiting the war on severe German terms.
- May 1918: Fighting begins between Czechoslovak Legions and Bolsheviks.
- July 1918: President Wilson sends a combination of 5000 US Army troops (American North Russia Expeditionary Force), and 8000 more troops (American Expeditionary Force Siberia) to Eastern reaches of Soviet Union to support Czechoslovak Legions, fight Bolsheviks, protect US and Allied interests, and reestablish Eastern Front.
- 1920 - 1924: Several hundred Americans move to Russia, establish socialist farm communes.
 - Two examples of communes named "Red Banner", and "Proletarian Life".
- March 7 - 17, 1921: The Kronstadt rebellion occurs.
 - A group of Russian sailors, soldiers, and civilians led by Stepan Petrichenko, a Russian revolutionary, organized the ultimately unsuccessful Kronstadt rebellion against the Bolsheviks during the later years of the Russian Civil War.
 - The Kronstadt rebellion was one reason why Communist Party decided to implement the New Economic Policy, which loosened government's control of economy.
- 1921 - 1922: The Russian Famine of 1921 AKA the Povolzhye famine occurs.
 - The Russian government allowed Maxim Gorky to ask foreign nations for aid.
- 1922: About 500 Americans form an industrial colony in western Siberia.
 - These Americans formed this colony in response to an article published by the radical journal The Liberator. The article asked Americans to develop the industries in Siberia to demonstrate the power of free workers.
 - The Americans also formed the colony in response to a letter from Lenin asking for American workers' help.
- October 1922: The Russian Civil War ends, resulting in the Bolshevik Red Army's triumph over the anti-Bolshevik White Army.
- January 21, 1924: Vladimir Lenin dies.
 - The cause of Lenin's death is widely suspected to have been syphilis.
 - Lenin wrote a Last Testament to be read after his death at a party congress. Party rulers suppressed its publication because it was anti-Stalin.
 - Max Eastman first publishes the Last Testament in 1925 in the United States.
- November 16, 1933: The US and Soviet Union establish diplomatic relations.
 - The US had cut diplomatic relations with Russia in December 1917, when the Bolshevik Party seized control of Russia. Franklin Roosevelt obtained the US presidency in 1933 and decided to re-establish diplomatic relations with Russia.

Appendix B: Sample Alternative Spellings of Russian Names

Alternative spellings of “Dostoevsky” *

“dostoevsky” OR “dostoyevsky” OR “dostoevskii” OR “dostoyevskii” OR “dostojevsky”
OR “dostojevskii” OR “dostoeffsky” OR “dostoyeffsky” OR “dostoeffskii” OR
“dostoyeffskii” OR “dostoieffsky” OR “dostoievsky” OR “dostoieffskii” OR
“dostoievskii” OR “dosteovsky” OR “dostoyefsky” OR “dostoievski” OR “dostoeffsky”
OR “dosteovskii” OR “dostoefsky” OR “dostoefskii” OR “dostojevsky” OR “dostojevskii”
OR “dostojevski” OR “dostoevski” OR “dosteovski” OR “dostoyevski” OR “dostojevski”
OR “dostojeffski” OR “dostoyeffski” OR “dostoeffski” OR “dostoieffski” OR
“dostoievski” OR “dostojevski” OR “dostoyefski” OR “dostoefski” OR “dostoiefski”

* Alternative spellings research conducted by Nick Slaughter of the Foreign Literatures in America project.

Appendix C: Scanning and OCR Guidelines



Russian Authors Initiative Manual

by Nick Slaughter, FLA Executive Editor – Russian Authors reception collection

Contents

I.	FLA Contact Information
II.	Cloud Storage – SugarSync
III.	Assignments a. Accessing Assignments b. Assignment Sheets
IV.	Scanning Materials a. File Formats b. File Naming c. Scanning Preparation d. Scanning Microfilm Documents e. Scanning Print Documents from McKeldin Periodical Stacks f. Scanning Print Documents from Off-site Storage g. Cropping Images h. Entering Bibliographic Data into Database Template i. Completing Scanning Assignments and Submitting Files
V.	Annotating Scans a. Annotation Questions Overview b. Annotation Procedure and Guidelines c. Completing Scanning Assignments and Submitting Files
VI.	OCR a. File Naming Conventions b. Coordinating OCR Tasks c. Using ABBYY FineReader d. Uploading Files for Processing to Master Archive and Database
VII.	File Storage

FLA Contact Information

Nick Slaughter FLA Executive Editor - Russian Authors reception collection Ph.D. Student, English Dept. Email: naslaughter.english@gmail.com Twitter: @naslaughter	Peter Mallios FLA Director Associate Professor, English Dept. Email: Mallios@umd.edu
--	--

Cloud Storage – SugarSync

Professor Mallios has purchased for the FLA project a 60GB cloud storage space through the service SugarSync; this space will be available to us as a central storage site and data transfer point. You should create a SugarSync account with the email address that you provided as your contact point for the project so that the FLA can give sharing permissions to each participant. SugarSync can be used either through its file manager application or through its web browser interface.

Five folders will be shared with participants:

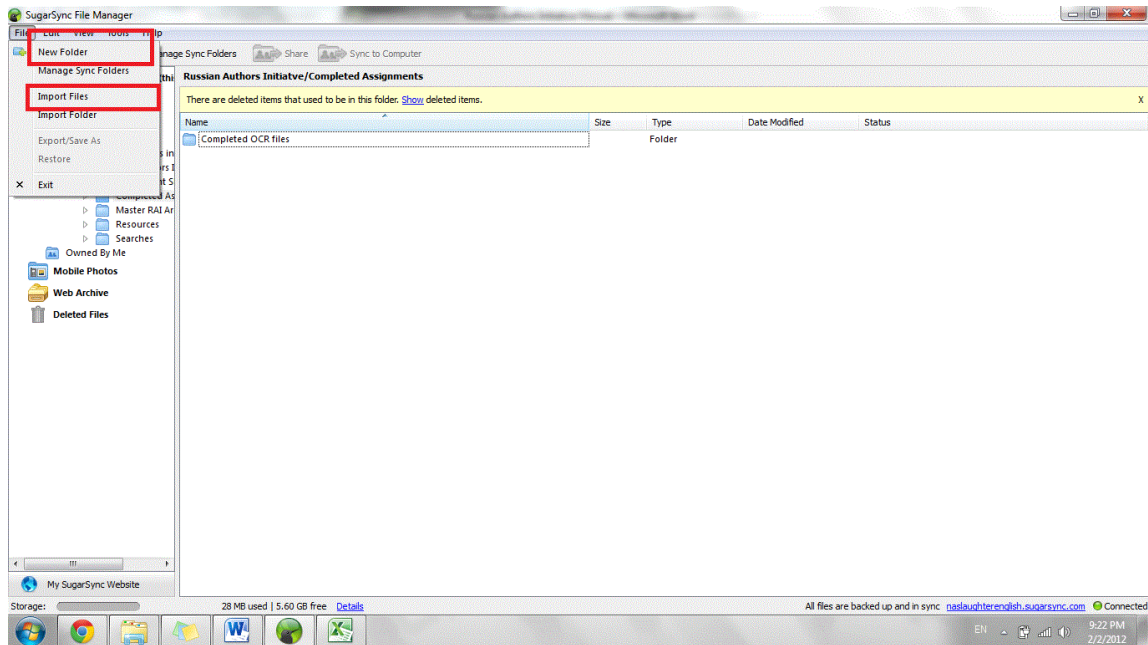
- Assignment Sheets – contains assignment information for participants
- Completed Assignments – space for participants to upload files to be integrated into the Master Archive
- Master RAI Archive – storage space for all files produced by the project as well as the Master Database
- Resources – contains various resources for project participants
- Searches – contains the results of bibliographic searches

Folders aside from “Assignment Sheets” and “Completed Assignments” folder will be read-only, meaning that participants will only be able to view original files and not edit them. Participants can, however, make copies of files in read-only folders as necessary. You should always make copies of files as you work with them rather than altering files stored in SugarSync

When uploading completed assignments, be sure to follow the instructions for how to do so in the section for each assignment. In order to upload files to the shared “Completed Assignments”, you can either use the file manager or the web browser interface.

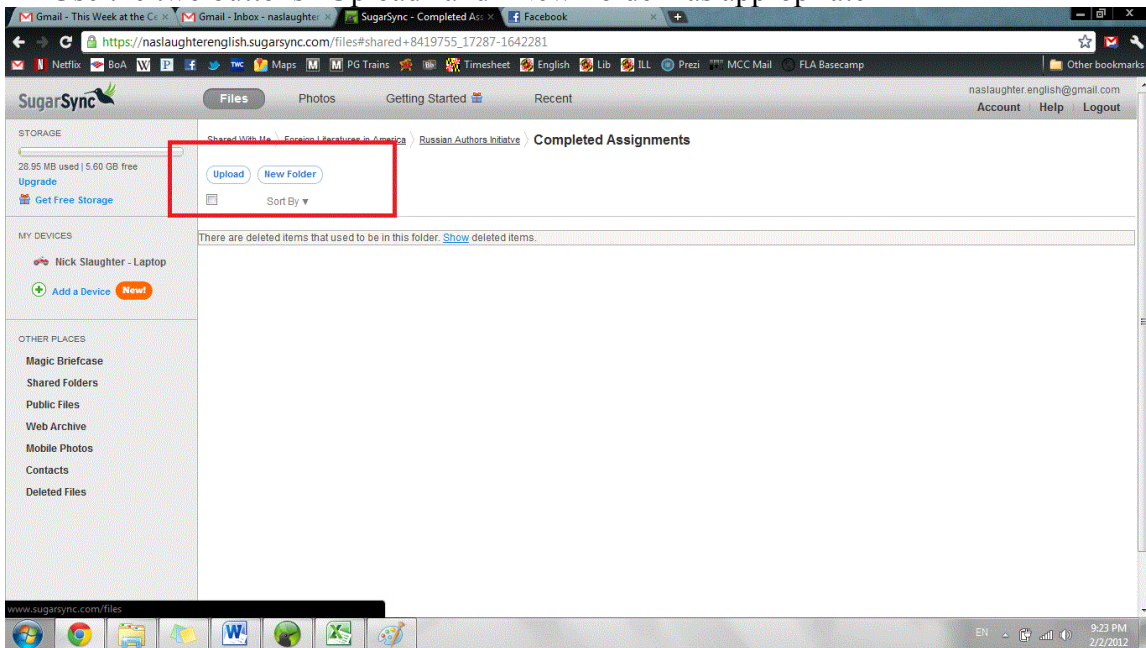
Using the file manager:

- Open the “Completed Assignments” folder
- Click on “File” drop down menu in order to create folders or import files as necessary



Using the web browser interface:

- Open the “Completed Assignments” folder
- Use the two buttons “Upload” and “New Folder” as appropriate



It is advisable to keep copies of all the files you produce for the project.

Assignments

This project will use assignment sheets and a tracking spreadsheet as its primary organizational tools. These will be generated by coordinating editors and available to all participants via SugarSync. In general, assignment sheets will be a method of both tracking the work of the project and making the project more efficient. There are three types of assignments, based on the three major stages of this project: scanning, annotating, and OCR. Assignments will generally not be “assigned” but chosen by participants; assignments do not need to be completed in any particular order, although a logical progression would likely be helpful.

Accessing Assignments

Assignment documents are stored inside the “Assignment Sheets” folder. To select an assignment, browse either the tracking sheet for available assignments or the assignment sheets themselves. Once you have selected an assignment, open the “Assignment Tracking sheet” and find the assignment that you have selected. Enter your name (or names if you are working in a group) in the participants column next to the assignment and the date on which you have selected the assignment. You will fill in the remaining columns once the assignment is completed. **Make sure that you save the spreadsheet before closing it.**

Assignment Sheets

Assignment sheets provide specific instructions for a task and should always be available to you when you are working on an assignment. Sheets are designed to help you track your own progress as well as provide space for you to make notes as necessary; since all of the research practices being undertaken here are still essentially experimental for both FLA and the Gemstone team, notes will help us refine our processes as well as communicate any discrepancies or nuances in our data among project participants.

You should fill out assignment sheets to be turned in once you have completed an assignment. You may do so either in hardcopy or electronically; coordinators will keep hardcopies of assignment sheets available in Tawes 3118. You may also download assignment sheets and print them out yourself. It would be wise to keep your assignment sheets with your copy of this manual.

For instructions on how to turn in assignment sheets and the relevant files, look to the end of the scanning, annotating, and OCR stages.

Scanning Materials

A scanning assignment generally requires participants to visit McKeldin library and use the resources there. There are generally two types of mediums to be scanned: microfilm and printed copies of periodicals. Of the print copies of periodicals, some are available in the periodical stacks of McKeldin, while others are stored off-site and need to be specifically requested. The scanning stage entails physically locating documents, copying them into a digital format, cropping images as necessary, and entering bibliographic data into a spreadsheet, and finally uploading files and spreadsheet to SugarSync to be combined with the master file archive and database. An FLA Database template file can be found in the “Assignment Sheets” in SugarSync.

Note: Always be willing to ask librarians for immediate assistance if you find yourself needing it, and always be willing to contact a coordinator or other FLA member if you have other questions.

File Formats

At this time, all scanned documents must be saved in a **.TIF format at resolution of 600 DPI**. TIF allows for the largest amount of data to be saved and is the only generally accepted format for OCR processes. .TIF is a very large format (generating large files), but you can save .TIF files with “loss-less” compression, either using “LZW” or “ZIP” compression that may be available as saving options in scanning or photoshop programs. If you have any questions, please consult a knowledgeable FLA member or MITH personnel as appropriate.

File Naming

All documents to be included in the FLA database should follow this naming convention:

fla-<research category>-<creator>-<document #>-<page #><version (optional)>.tif

Assignment sheets will provide instructions for file naming in regards to “research category,” but this field is relatively simple; for example, the research category for files dealing with the reception of Tolstoy would be “tolstoy”. Research categories will be tracked in the FLA Conventions and Categories database available in the “Resources” folder in SugarSync.

The “creator” field is the designated identifier of an FLA participant, using his or her three initials and a numeral. Identifiers will be assigned by coordinators and saved in the FLA Conventions and Categories database.

- Ex: “nas1” for Nick Slaughter
- Ex: “jjw1” for Jennie Wellman

The “document #” field enumerates documents in the format <0000> . Each document for a research category receives an arbitrary four-digit number to help group files together appropriately. Conventionally, files should be enumerated beginning with “0001”; participants do not need to coordinate their document numbers with other participants, meaning that Nick Slaughter and Jennie Wellman could respectively have files “fla-dost-nas1-0001-001.tif” and “fla-dost-jjw1-0001-001.tif” and these file names would not conflict with each other.

The “page #” field enumerates the page of a document that a particular file contains, in the format <000>. The first page of a document should be numbered “001”; the second, “002”, and so forth. This convention is for tracking the order of document pages rather than the actual page numbers within a document.

The “version (*optional*)” field is optional and intended to differentiate between separate versions of the same scan or page of source material. The main distinction between versions is a “cropped” version and a “full-page” version; not all original documents are the sole article on the page of a source, so for FLA purposes the targeted article needs to be cut out or cropped from the full page. This must be done for effective OCR to take place. However, FLA values having the full-page version available as well; whenever possible, both versions should be saved. At this time, **the “cropped” version of a file that is more or less OCR ready should be version “A,” with other versions being labeled as “B,” “C,” etc. as necessary.**

File Name Examples:

fla-dost-nas1-0026-002.tif

fla-tolstoy-plm1-0084-011.tif

fla-chekhov-jjw1-0258-003A.tif

Scanning Preparation

In order to begin creating effective digital scans of documents, you will need to have the following with you:

- a large flashdrive or other portable digital storage device
- an electronic copy of the FLA Database template downloaded from SugarSync
- your folder with manual and assignment sheet(s)
- a soft cloth and cleaning solution (a basic glass cleaner like windex is sufficient)

You may also wish to bring a ruler with you to help scan print documents.

Scanning Microfilm Documents

Microfilm reels are available at the back of the first floor of McKeldin to use in the various microfilm viewers/scanners in the same area. Coordinators will normally provide the relevant call numbers and reel numbers that you need to scan, but microfilm can be located using the library system’s electronic catalog. **Note:** Not all microfilm reel boxes are specifically labeled; you may have to use trial and error to locate a specific reel. Once you determine what an unlabeled box contains, it is helpful to use a pencil to write on the box what periodical and volumes are contained within for future users, including yourself.

Although the library requests that you only use five reels of microfilm at one time, depending on the scope of your particular assignment, it is advisable to simply collect all the reels you need or plan to use in a single sitting at once. When you have acquired your microfilm, select a scanning station and log into the computer using your UMD Directory ID and password.

The scanner software is called “PowerScan 2000”; double-click the icon that should be on the computer desktop and when prompted choose “35mm Microfilm” as your medium. Before you load a microfilm reel into the scanner, click the “File” tab at the bottom of the screen and set the scanning resolution to **600 DPI**. **Note:** If you close PowerScan and reopen it, the program will

reset to a default scanning resolution of 300 DPI, so be sure to check the scanning resolution every time you open the program.

Before loading a microfilm reel, be sure to check the glass tray for marks and debris. Use your cleaning solution as necessary to clean the glass, and wipe down the glass thoroughly to remove any dust or other small particles that could cause a scan to become unreadable by OCR. Even as you are scrolling through a reel of microfilm, particles can fall out of the reel and dirty the glass; carefully brush away as many particles as you can so that you can make a clean scan.

An on-screen diagram should appear to show you how to properly load the microfilm reel. In order to load the reel, you must pull the glass tray out towards you; the tray will open for you to pull the microfilm through and attach it to second spool. You will usually need to tell the microfilm viewer to rotate 90 degrees so that the microfilm appears right side up on the monitor; you can find the button to rotate on the bottom of the screen under one of the tabs. There are three speed settings you can use to scroll through the reel: super-fast-forward and –rewind by pulling the glass tray all the way out and using the on screen buttons; regular fast-forward and rewind when the glass tray is pushed all the way in; and page by page scrolling. The more often you use the microfilm scanners, the more adept you will become at using these controls to quickly find the documents you want.

Note: If you cannot find a particular option or command in PowerScan 2000 listed in this manual, then it might be disabled; ask a librarian to enable it for you.

Once you have located the document you are seeking, use the “Auto Adjust” button (under either the “Home” or “Adjust” tabs) for the scanner to automatically correct its focus and lighting for that particular page. You might also need to use the manual rotate buttons at the bottom of the screen under one of the tabs to straighten the image; images do not have to be perfectly straight, but some straightening certainly creates a more pleasant product. Adjust the green cropping box on the screen to scan the document, page by page; be sure to include page numbers and other essential information in your scans, but leave out wide margins and other blank space when possible.

To save a scan, click on the “File” tab and use the button “Scan to Drive #1”. This button should open a “Save As” window through you can save your file to any location on the computer. Remember to save all images in .TIF format; the scanning software also allows you to save a .TIF file with LZW compression, which saves space. **It is much quicker to save scans to the computer’s hard drive initially than to a flash drive. You should find a temporary folder to use on the hard drive (C:/) to save scans before transferring files to your external drive.** For articles that share a page with other articles, be sure to scan the full page as well as crop the targeted article as best as you can. Remember to name files appropriately, and record file names and other relevant bibliographic information on your assignment sheet. It is advisable to periodically copy files to your flash drive since waiting until you have finishing scanning will leave you to wait for a bulk of files to transfer.

Once you have scanned the full document, be sure to check and confirm or correct the bibliographic data provided on your assignment sheet. If there are any title discrepancies, record

all apparent article titles that appear on the actual pages on the source material. Try to find the cover or the front page of the source you are browsing in order to confirm the issue number, date, publisher, publisher location, etc. of a source.

To quickly rewind a reel, pull the glass tray fully out and double click the rewind button.

Scanning Print Documents from McKeldin Periodical Stacks

Some printed copies of periodicals are available in McKeldin's Periodical Stacks. A coordinator will normally provide call numbers of printed periodicals on each assignment sheet, and you will locate the periodicals to scan. Printed copies are large and bulky, so you may not be able to carry more than three or four at one time. It may be advisable to divide labor between two participants at once.

Note: You may discover that some periodical volumes are in poor condition and will be further damaged if you try to scan them on a conventional scanner. If this is the case, make a notation to communicate this fact to a coordinator. Hornbake library has overhead scanners which may facilitate easier scanning of damaged materials.

Note: Handheld scanners, digital cameras, etc. may be able to be used to make this scanning process more effective, but all scanning methods need to be approved as appropriate by the FLA Board in consultation with MITH personnel.

Full-sized scanners are available in McKeldin; one is located at the front of the library by the reference computers, to the left of the main entrance when you first enter; the others are located on the second floor. The following instructions are for use with scanners connected to PCs; detailed instructions are not available for scanners attached to Macs, but the principles should be the same or similar.

When you have acquired your print periodicals, select a scanning station and log into the computer using your UMD Directory ID and password. The scanning software is titled "EPSON Scan" and should be available on the desktop. Select the following parameters from the initial options:

- Grayscale
- 600 DPI resolution
- DO NOT enhance text

You must also define the orientation and dimensions of the area to be scanned; it is permissible and sometimes recommended to scan two pages as one file; each individual scanner can determine when this is a good idea or not. A ruler comes in handy when defining a custom scanning area, but you should always err on scanning a slightly larger area than the periodical, cropping margins later.

Once you've selected your initial options, open the scanner bed and clean the glass as necessary, wiping away particles with your cloth. You may choose to leave the scanner bed open or closed; in many cases it will be difficult to close the scanner bed, and you can later remove any unwanted margins in the scan by cropping them out.

After you've set up the item to be scanned and selected from your initial options, clicking the "Scan" button will open a second set of options. Be sure to select **.TIF** as your file format. You also need to choose a location in which your files will be saved. **It is much quicker to save scans to the computer's hard drive initially than to a flash drive. You should find a temporary folder to use on the hard drive (C:/) to save scans before transferring files to your external drive.** It is advisable to periodical copy files to your flash drive since waiting until you have finishing scanning will leave you to wait for a bulk of files to transfer. The scanner software offers an automatic method for naming files; you may use this method, inputting your own prefix, or you can give files a temporary name to be renamed later.

When scanning, be sure to hold the periodical firmly against the glass; a flatter scan will produce a clearer, more legible scan for OCR. Of course, be careful with how you handle the volumes so as not to damage them.

Note: Older volumes that are deteriorating will likely leave paper particles on the glass between scans. Be sure to check periodically to see if the glass needs to be wiped down.

Once you have scanned the full document, be sure to check and confirm or correct the bibliographic data provided on your assignment sheet. If there are any title discrepancies, record all apparent article titles that appear on the actual pages on the source material. Try to find the cover or the front page of the source you are browsing in order to confirm the issue number, date, publisher, publisher location, etc. of a source.

Scanning Print Documents from Off-site Storage

A large number of printed copies of periodicals are actually stored off-site from McKeldin library. These volumes can be requested through this web-page, which will be listed on every relevant assignment sheet: <http://www.lib.umd.edu/PUBSERV/jnlrecall.html>

After a day or two, the library will transfer the requested volumes to McKeldin to be kept temporarily behind the circulation desk. Thus, this type of assignment requires some more planning ahead of time, but otherwise the scanning the process the same as for other print documents.

Cropping Images

This step can be done with various photoshop software like Adobe Photoshop or GIMP (available for free online). When resaving cropped .TIF files:

- be sure to follow the "version" convention listed in the file naming convention
- be sure to resave .TIF files with loss-less compression like LZW or ZIP

DO NOT alter images in any other way aside from cropping; any other changes to images are likely to alter the underlying data in a file and could corrupt the OCR process. Any sort of alterations aside from cropping to final scans will be done at the OCR stage as necessary.

Entering Bibliographic Data into a Database Template

Once you've acquired all the documents listed on your assignment sheet (or as you collect each individual document), you will enter the basic bibliographic data into an FLA Database template that you have downloaded from SugarSync. Each and every file that you have created and plan

to submit to SugarSync should be represented in an individual row. **Save the spreadsheet that you create as “Scanning #1 - <your name>” so that spreadsheets can be tracked by both assignment and participant once uploaded to SugarSync.**

You should enter information for the following fields:

- File Name
- Page #(s)
- Pages in Document
- Main Title
- Sub Title
- Alt Title
- Descriptive Title
- Author
- Placement in Publication
- Publication
- Volume
- Issue/Number
- Date (Month.Day/Season)
- Year
- Publisher
- Publisher Location
- Date Acquired

Every field can be copied and pasted for all the files in a single document except, of course, “File Name” and “Page #(s)”.

“File Name” – this should be an exact transcription of a file name, including the .tif extension.

Advanced: You may use the file “Name generator” in the “Resources” folder to automatically generate a text file that lists all the file names in a given folder. To do so, you must copy the “Name generator” file into the folder with your scans and double click the file. A text file named “list” should appear in the folder that contains all your file names in text that can be copied and pasted into your spreadsheet.

“Page #(s)” – enter in this field the actual page numbers that appear in an individual scan. For more than one page, enter the pages as <000-000>, never abbreviating the page numbers.

- Good example: “134-135”
- Bad example: “134-5”

“Pages in Document” – enter in this field the full range of pages from the document, never abbreviating the page numbers (see above example).

“Main Title” / “Sub Title” / “Alt Title” – these fields are divided into three because periodicals do not have a systematic, consistent method for titling articles, and not all articles that you will be locating will have clearly defined titles. Generally speaking, the most prominent title of a document should be listed as the “Main Title”; **the bibliographic references provided on**

assignment sheets will not always provide an accurate “Main Title”, so be sure to confirm this information. Include in the “Sub Title” and “Alt Title” fields any additional titles that may be present. If these fields are not applicable to an article, enter a null value or leave that field blank.

“Descriptive Title” – this field is for capturing important information about an article that is not represented in the actual main title or sub titles of an article. Book reviews commonly deserve a “Descriptive Title”. For example, an article simply titled “Dostoevski” was really a book review, so in the “Descriptive Title” field was entered “Review of The Complete Works of Dostoevski, translated by Constance Garnett. New York: The Macmillan Co.” This field is a descriptive category that does not follow strict conventions; please enter information here that you think is important to understanding what the article is about. If you are uncertain about the validity of any of your annotations, communicate this with a coordinator. If this field is not applicable to an article, enter a null value or leave this field blank.

“Author” – enter in this field the author’s name as it appears in the article, following a convention of <Last Name, First Name>. For authors listed only by their initials, enter the initials as they appear in the article. If no author is listed, enter a null value or leave this field blank.

“Placement in Publication” – like the “Descriptive Title” field, this field is a descriptive category without strict conventions. To the best of your ability, please describe where the article is situated within its source. If you cannot provide an informative description, leave this field blank.

- Example: “First article of issue”
- Example: “In the middle of the Leading Articles section”

“Publication” – enter in this field the name of the publication in which you found the targeted article.

“Volume” – enter in this field the volume of the source in which you found the targeted article; almost always will this information be correct in the bibliographic reference provided in an assignment sheet.

“Issue/Number” – enter in this field the “issue” or “number” of a volume in which you found the targeted article. This information is not always provided in bibliographic references, so you should always check the cover or first page of an issue for this information. Not all periodicals will have an issue or number listed; if so, leave this field blank.

“Date (Month.Day/Season)” – enter in this field the date or season on which this article was published. For dates, follow a convention of <00.00> for <Month.Day>.

- Some issues will only list a season as a date of publication; enter the season in this field as it is printed in the issue.
 - Example: “Autumn”
 - Example: “Spring”

- Some issues will only list a month as a date of publication; enter the month as a two-digit numerical value and leave the day as “00”.
 - Example: February – “02.00”
 - Example: December – “12.00”
- Examples of how to record specifically listed dates:
 - March 31 – “03.31”
 - July 2 – “07.02”
 - October 13 – “10.13”
- If no date whatsoever is listed, leave this field blank.

“Year” – enter in this field the year in which the document was published; almost always will this information be correct in the bibliographic reference provided in an assignment sheet.

“Publisher” – enter in this field the name of the publishing company for the source, which can be often found on the cover or first page of a periodical. If this information is not available, enter a null value or leave this field blank.

“Publisher Location” - enter in this field the location of the publishing company for the source, which can be often found on the cover or first page of a periodical. Enter as <City, ST> or <City, Country> as applicable. If this information is not available, enter a null value or leave this field blank.

“Date Acquired” – enter in this field the date <mm/dd/yyyy> on which you scanned each file.

Leave all other fields blank.

Completing Scanning Assignments and Submitting Files

After you have completed each part of an assignment, create a folder inside the “Completed Assignments” folder in SugarSync. Label the folder with the title of the assignment and your last name. Example: “Scanning #1 – Slaughter”. Copy your finalized scans and database file into this folder. If you also are submitting an electronic copy of your completed assignment sheet, copy it into this folder; if you are submitting a completed hardcopy of the assignment sheet, please turn it in at Tawes 3118. Then open the Assignment tracking spreadsheet and enter the date you completed the assignment as well as how many hours you think the assignment took to complete.

Once you have submitted all the materials for an assignment, a coordinating editor will review your notes and create a Unique Identifier (UUID) for each document as well as complete the “Source”, “Location Acquired”, “Medium Acquired”, and “American or British”, etc. fields. Then your scans and database entries will be combined with the master archive and database, ready to be processed by the annotating and OCR teams.

Annotating Scans
THIS SECTION IS OMITTED
A REVISED VERSION CAN BE FOUND IN APPENDIX #.

OCR

The process of transforming digital images into digital text files will take place in Tawes 3118, an office reserved for FLA use. The OCR software package we're using is called ABBYY FineReader.

At this time, we are producing **two** “output” files of the OCR process: a plain .txt **UNCORRECTED** transcription and a .DJVU file that store character coordinates. In addition, we will be saving a reference file in the FineReader document format.

File Naming Conventions

Our tentative file naming convention for OCR output files is to add an appropriate prefix to the file name of the relevant image. The prefixes are:

- “OCR-A-” – uncorrected plain .txt files
- “OCR-B-” – uncorrected .DJVU files
- “OCR-C-” – uncorrected FineReader Document files

Example: OCR-A-fla-dost-nas1-0003-001.txt

Coordinating OCR Tasks

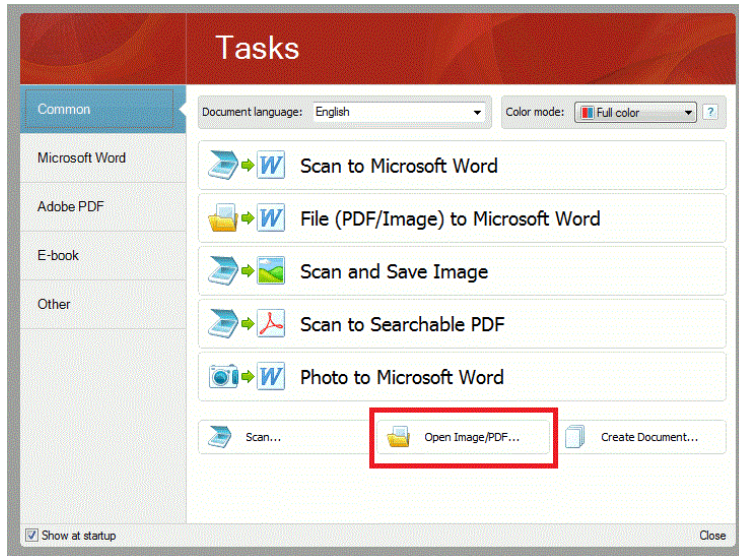
OCR tasks will be coordinated through a checklist printed and stored in Tawes 3118. Once you have created an output file for an image file, write your initials in the appropriate box on the checklist. The checklist also provides a section for notes; please record there any special information or difficulties you encountered during the OCR process so that these issues can be addressed.

Files for the Russian Authors Initiative are saved under “Documents/Foreign Literatures in America/Russian Authors Initiative”.

To access image files, you can either use images stored on the PC hard drive or you can access images stored in SugarSync. At this time, you should only scan “A” versions of files—images that have been cropped of additional material on the page. **Note:** All cropped images may not be correctly labeled as “A”, so be sure to double check between “A” and “B” images. When saving output files, you should initially save them to the PC hard drive in the appropriate folders inside the “OCR Workspace” folder on the computer’s desktop and then copy them into SugarSync into the appropriate folders inside the “Completed OCR files” folder inside the “Completed Assignments” folder so that they can be processed and added to the master archive and database.

Using ABBYY FineReader

After opening ABBYY FineReader using the desktop shortcut, you will be prompted with the following window:

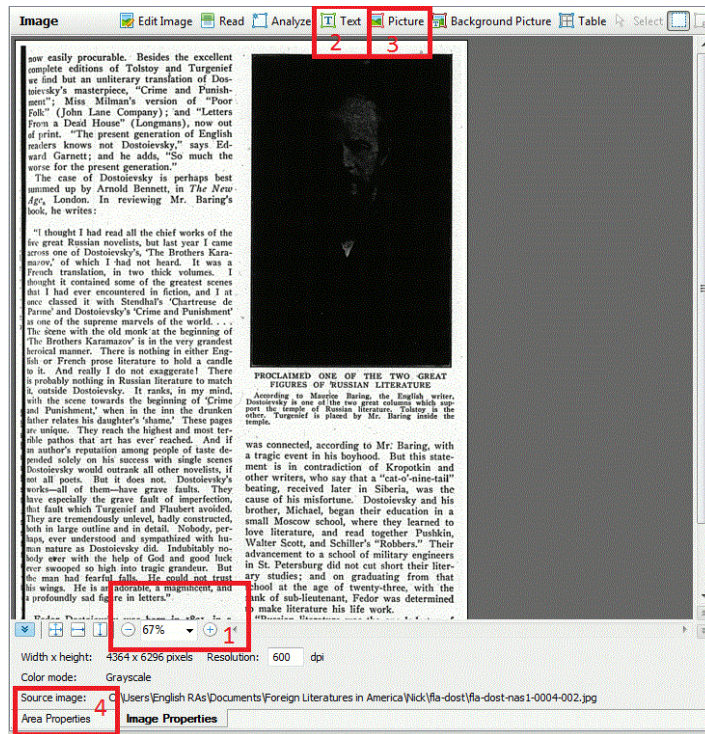


Click the button for “Open Image/PDF” (highlighted) and select the file you are going to OCR. At this time, we will not alter the default options when opening a file.

Note for “inverted” black and white images: See the *Appendix* section of at the end of the OCR section of this manual.

Step 1. – Drawing “Areas”

In order for ABBYY to run its OCR process, you must select the sections of an image for it to read and analyze. This is known as drawing an area. There are two main types of areas we will be using: Text areas and Picture areas.



When you first open an image, you should change the magnification setting (highlighted box 1) so that you can see the full image.

Next, begin drawing text areas by clicking the Text button (highlighted box 2) and dragging rectangular boxes around columns of text. Each individual column of text must be contained in a single area; do not draw areas around multiple columns, as ABBYY will not read this text appropriately. You should draw text areas as tightly as possible around columns of text because ABBYY might interpret extraneous marks on the image as typographical characters.

- **Important:** The format of your output files depends on the “order” of the areas you draw; you can alter the order of your areas after you have drawn them. You should first draw text areas around the text **that constitutes the main text of the document** as opposed to footnotes, image captions, or other contingent text on a page. At this time, you should scan this contingent text last, so that it appears at the end of your plain .txt output files.
- For bodies of text that are not shaped in simple rectangles, you can use “Add Area” or “Cut Area” tools that appear next to the mouse if you select a text area.

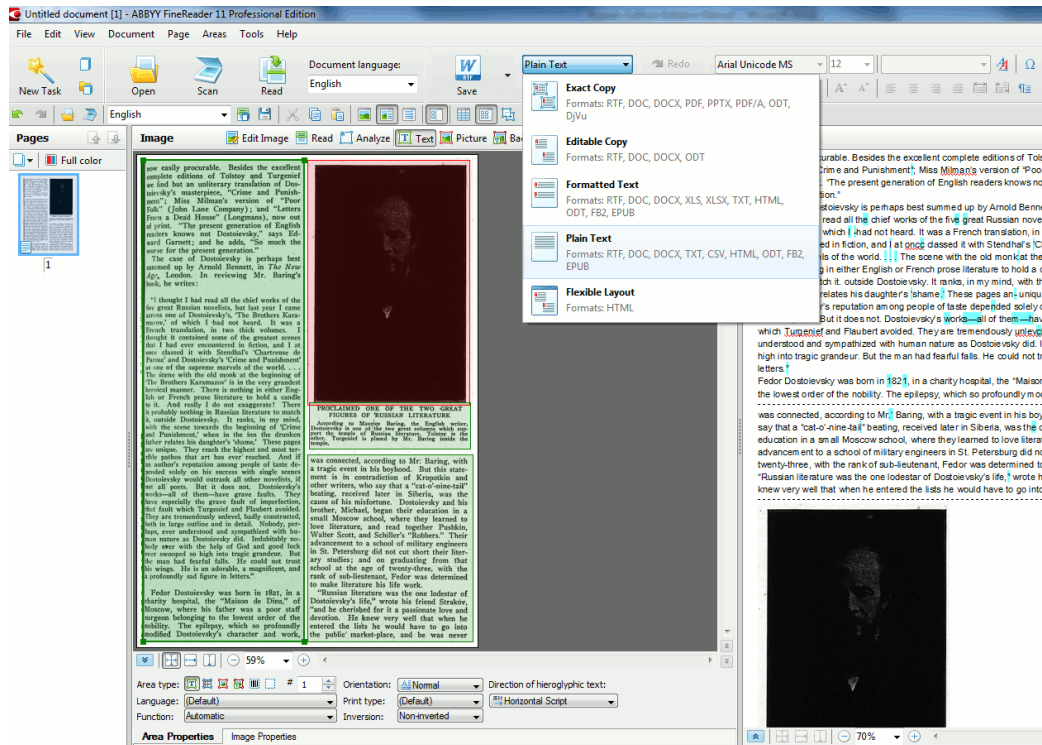
Lastly, draw picture areas as needed by clicking on the Picture button (highlighted box 3) and dragging rectangular boxes around pictures. For pictures with captions, you should reorder the Text area that includes a caption so that it appears after the image.

To delete an area, left click on it and press the Delete key on your keyboard. You can use the “Undo” and “Redo” functions in the edit menu to help you edit your areas.

From time to time, you will need to alter the order of your areas. To do so, click on the Area properties tab (highlighted box 4) once you have drawn your areas.



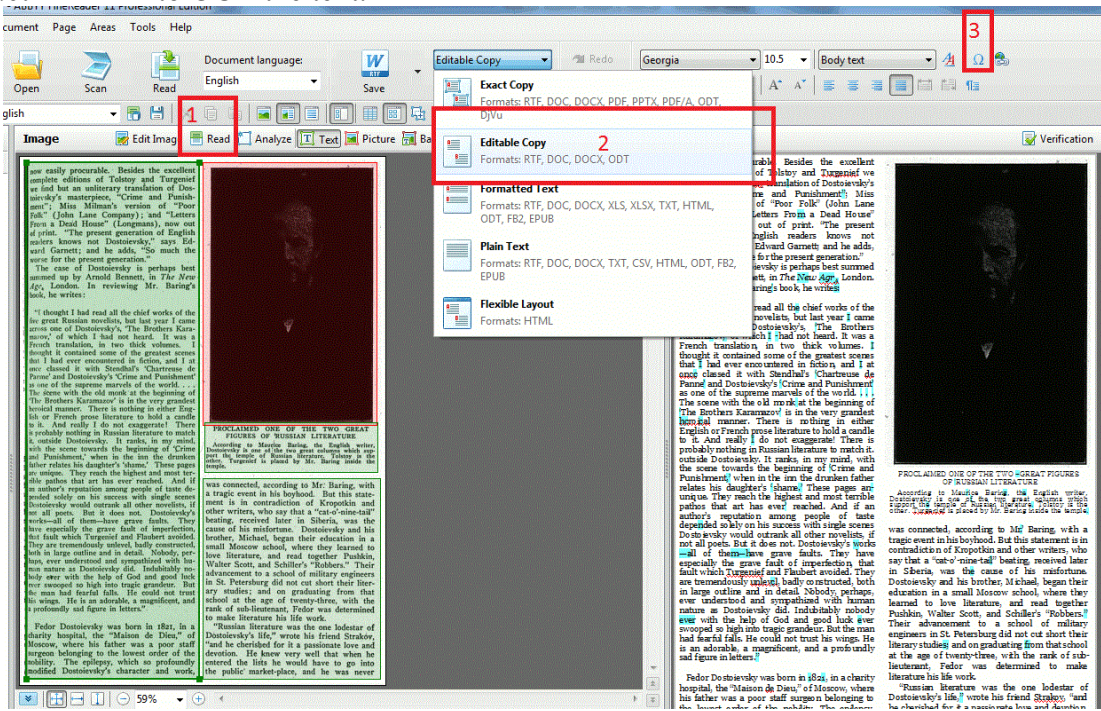
Once you have selected an area by clicking on it, you can alter the order of that area by changing the value in the “Area #” box (highlighted above). To change the value, enter a new one and be sure to press ENTER for the value to be saved. You can quickly check the order of your areas by seeing the order of their output in the “Plain Text” view after you have instructed ABBYY to “Read” your text areas (see below image).



DUE TO TIME CONSTRAINTS, TEAM POLITIC ELIMINATED STEP TWO.

Step 2. Reading and Editing Text

Once you have drawn your areas, click the “Read” button (see highlighted box 1 below) to instruct ABBYY to OCR the text.



There are five output views that ABBYY offers, but we are mainly concerned with the following three:

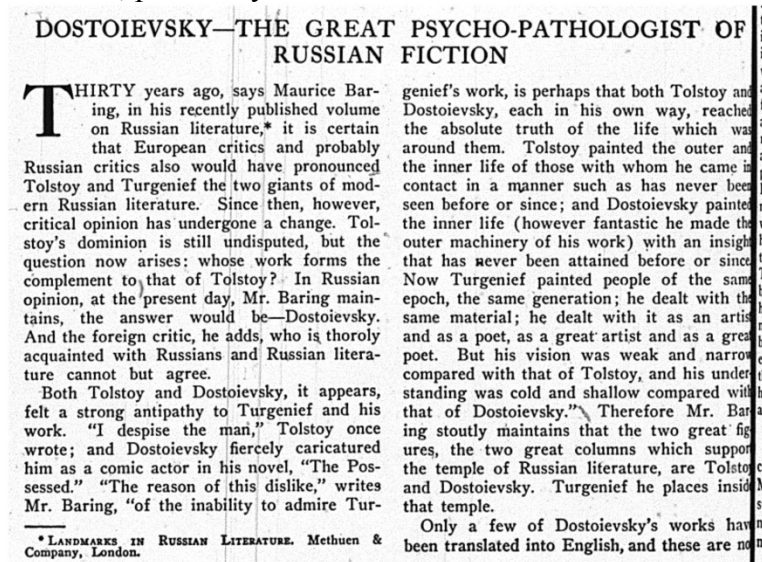
- “Exact Copy” produces an uneditable, precise as possible facsimile of the text on the original image
- “Editable Copy” produces an editable, slightly less precise facsimile of the text on the original image
- “Plain Text” produces a plain text format very similar to that of a .txt file

The best view to use for editing is the “Editable Copy” view (highlighted box 2). The blue highlights and red marks in the output view indicate errors that ABBYY thinks it might have made, but not all highlighted areas are actually mistakes. **Important:** Like spell checking in any other word processing program, ABBYY will make mistakes or mark typos in the original image as incorrect; you should not change any original spellings that are present in the original image. **This means that you must read through the entire OCR output to check for accuracy against the original image.**

If you come across special characters in the original that are not reproduced accurately in the OCR output, you can find special characters using the symbol tool (highlighted box 3).

Important: If you come across names that are hyphenated across columns or pages, edit the OCR output text so that the full name appears together.

- Example: “Turgenieff” is split between two columns and ABBYY reads it as “Tur-” and “-genieff”; this results in a later text analysis not recognizing the split name as “Turgenieff”. Thus, you should edit the output text to have “Turgenieff” appear as one word in a single column, preferably the first column.

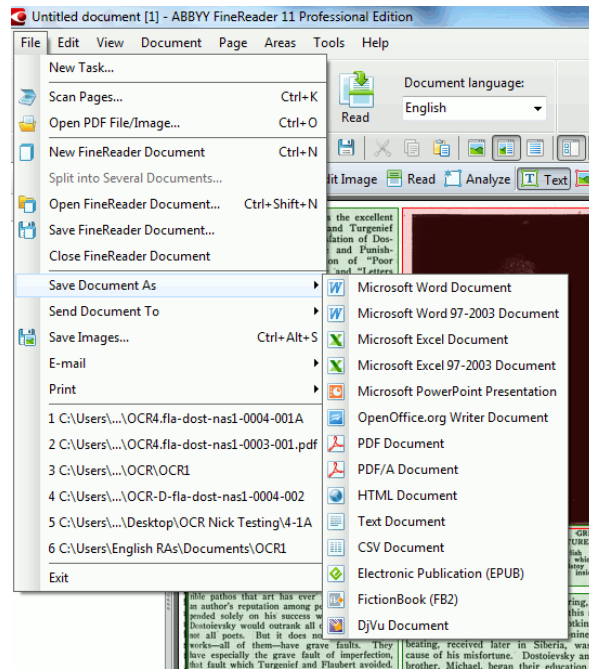


Step 3. Saving Output Files

Once you have finished editing the OCR output, you need to save it in the four separate file formats:

- Plain .txt file (prefix “OCR-A-”)
- .DJVU file (prefix “OCR-B-”)
- FineReader document format (prefix “OCR-C-”)

You can save in each of these formats by going under the “File” drop down menu and selecting “Save Document As” (see image below). At this time, you should save each different output format into the designated folder in the “OCR Workspace” folder on the computer’s desktop. You should leave the options for each format at their saved default settings.



Uploading Files for Processing to Master Archive and Database

Once you have generated OCR output files, you should copy them into SugarSync into the appropriate folders inside the “Completed OCR” folder inside the “Completed Assignments” folder so that they can be processed and added to the master archive and database. Be sure to mark on the checklist when you have completed files, and make any notes as necessary.

File Storage

All files for the Russian Authors Initiative will primarily stored in SugarSync while our team is doing work. The computer in Tawes 3118 and Nick’s personal desktop computer will automatically back-up any files uploaded through SugarSync; Nick will also copy files directly to his laptop. Participants are welcomed to volunteer to create additional back-up sites, and participants are always encouraged to keep copies of any files they produce.

Appendix D: Sample Annotation Question Evolution

Current Sample Annotation Question

4. **Sentiment Analysis: Principal Author as Subject of Debate.** (Y/N) Does this article contain any explicit reference to the literary author(s) it principally concerns as a subject of debate, either because interpretations of that literary author's meaning are explicitly disputed, or because opposing positive and negative opinions of an author are explicitly referenced?

Original Sample Annotation Question

4. **Sentiment Analysis: All Opinions Expressed in the Article.** [This question concerns *all* opinions expressed in the article concerning the literary writers in question—whether they express the article's own point of view or other perspectives quoted and referenced in the article.] Which of the following ratings comes closest to the *entire field* of opinions quoted or mentioned in this article concerning each of the literary authors the article principally concerns? *Note: this question should be answered separately for each author named in question 1.*

- 2 – **A Positive Opinion:** a generally or ultimately positive opinion as an overall matter
- 1 – **A Mixed or Unclear Opinion:** such that it is not possible to say whether the article's overall opinion of an author is positive or negative
- 0 – **A Negative Opinion:** a generally or ultimately negative opinion as an overall matter
- X – **Neutral:** This article is not evaluative: it does not express opinions about the author(s) in question, but is rather strictly and neutrally factual

Appendix E: Annotation Questions and Guidelines

- **Author (or authors) of principal concern in article.** What literary author or authors, if any, is this article primarily about?
 - *Spelling:*
 - Be sure to spell any names given in answer to this question *as accurately as possible*, exactly reproducing how the name is spelled in this article. (Spellings will differ between articles: we want to capture the differences.)
 - Include the *fullest version* of the author’s name included in the article: i.e., include an author’s first and/or middle names and/or initials if these names are included at any point in the article.
 - *Individuals:* Only literary authors named by *personal name* (i.e., not anonymous figures or those referenced only by job title) and who are *persons* (i.e., not publications) count as “authors” for purposes of this question.
 - “Literary author” means an author of fiction, poetry, plays, or related forms of creative writing. This applies whether the author is being invoked in his or her capacity as a literary writer or not. *Academic professors, literary critics, and journalistic and other commentators on literature do not fall into this category, unless they have significant literary accomplishments of their own.*
 - An author is of “principal” or “primary” concern in an article when an author is a major, continual, or focal concern that runs and receives explicit mention *throughout* an article as part of its general field of concerns, not just in discrete or severable paragraphs of it.
 - Some more rules of thumb on identifying whether an author is a “primary” or “principal” concern in an article:
 - if a literary author’s name is **included in the article’s title**, it is likely that s/he should be included in the answer to this question
 - if there is a large disproportion between the number of times different authors are mentioned or referred to, this is a good indicator that those mentioned less should likely **not** be included in the answer to this question
 - if the excising of relatively few paragraphs from this article would result in the elimination of reference to an author, that author should generally **not** be included in the answer to this question
 - As a general matter, **construe answers to this question narrowly**: only an author (or authors) comprising the main and consistent focus of an article should be included—although articles whose explicit focus is evenly to compare two (or more) authors throughout may be described as having multiple “principal” authors
- **Sentiment Analysis 1: the Opinion of the Article Writer.** Which of the following ratings comes closest to the *article writer’s* expressed opinion of the literary author(s) this article principally concerns? [Note: this question concerns the opinion ultimately taken by the *article writer him/herself* on the literary authors question. This is so even

though the article writer may quote or reference opposing opinions along the way.] *This question should be answered separately for each author named in question 1.*

- 2 – **A Positive Opinion:** a generally or ultimately positive opinion as an overall matter.
- 0 – **A Negative Opinion:** a generally or ultimately negative opinion as an overall matter.
- U – **A Mixed or Unclear Opinion, or No Opinion Offered:** it is not possible to say whether the writer’s overall opinion of an author is either positive or negative because the writer’s opinions are mixed, unclear, or not offered at all.
- **Sentiment Analysis 2: Uncertainty of Article Writer’s Opinion.** If the answer to Question 2 is “U,” answer the following question; if not skip it. Which of the following ratings comes closest to describing why the article writer’s opinion of a principal literary author is unclear? *This question should be answered separately for each author named in question 1.*
 - 1 – **A Mixed or Unclear Opinion:** the article writer either expresses mixed opinions about the literary author, or does not make clear how the opinions, judgments, or values s/he holds clearly relates to the literary author
 - X – **Straight Factual Account:** this is not an article in which the article writer’s personality, opinions, judgments, are in evidence; the article writer assumes the position of the “straight,” factual, objective newspaper reporter; the article writer’s stance is *neutral* with respect to his/her own opinions and values, not evaluative.
- **Sentiment Analysis 3: Principal Author as Subject of Debate.** (Y/N) Does this article contain any explicit reference to the literary author(s) it principally concerns as a subject of debate, either because interpretations of that literary author’s meaning are explicitly disputed, or because opposing positive and negative opinions of an author are explicitly referenced?
- **Books mentioned?** (Y/N). Does this article explicitly mention by title any specific books, poems, or texts written by any literary author it is principally about? *Note: this question should be answered separately for each author named in question 1.*
- **National identification.** (Y/N) Does this article specifically identify the nationality of any literary author it is principally about? *Note: this question should be answered separately for each author named in question 1.*
- **Style or literary artistry as issue.** (Y/N) With respect to any literary author this article is principally about, is the author explicitly described in terms of “art” or as an “artist” or in terms of his or her “artistic” vision, or is at least one paragraph of the article devoted to the style (not the content) of his or her writing? (A “yes” answer to any part of this question means a YES answer to the question as a whole.) *Note: this question should be answered separately for each author named in question 1.*
- **Foreign Place Names.** (Y/N) Are there any non-U.S. place names mentioned in this

article?

- **Gender of Article Writer.** Use the following scale to identify the apparent gender of the writer of this article (i.e., *not* the gender of the literary figure(s) in question, but the gender of the article writer who is writing about the literary figure(s)):
 - M – Male
 - F – Female
 - U – Unclear (i.e., because name is ambiguous or initials are used; the article is unsigned; or for another reason)
- **Gender as Issue.** (Y/N) Is gender ever explicitly discussed as an issue in this article?
 - Note: The fact that a character or author discussed in the article is a man or woman is not sufficient to constitute a Yes answer to this question; there needs to be some explicit attention drawn to gender as a matter of significance—(if only in a single phrase)—or reflection on or significance attributed to the categories of “man” or “woman,” “masculine” or “feminine,” or other gender ideas.
- **Race as Issue.** (Y/N) Is race ever explicitly raised as an issue in this article?
 - Note: this question should be answered “Yes” only if: (i) the article explicitly uses the term “race” (or some direct variant on it: “racial,” “racism,” etc.); (ii) there is explicit discussion about general ideas of race; or (iii) one of the following radicalized categories is explicitly invoked: black or African; white or Aryan or Caucasian; Slavic; Jewish or Hebrew.
- **Socioeconomic class as issue.** (Y/N) Does socioeconomic class receive explicit discussion in this article?
 - Note: Any explicit mention of social class (for example, “aristocratic,” “peasant,” “the poor,” “Count,” “prince”) will qualify as a YES answer to this question. (Czar, however, as a state figure, does not alone qualify.)
- **Religion as Issue.** (Y/N) Does religion receive explicit discussion in this article?
- **Radical Politics as issue.** (Y/N) Do any radical political movements including anarchism, nihilism, bolshevism, socialism, or communism receive explicit mention in this article?
- **America/West invoked as a point of similarity with Russia.** (Y/N) Does this article make any specific and explicit claims that Russia shares any quality in common with the U.S., “the West,” or any of the countries, cultures, and/or literatures of Western Europe?
- **America/West invoked as point of contrast with Russia.** (Y/N) Does this article draw any specific and explicit contrasts between Russia or anything Russian and any qualities or aspects of the U.S., “the West,” or any of the countries, cultures, and/or literatures Western Europe?

Appendix F: Downloading WEKA and Generating an ARFF File

Downloading and Optimizing WEKA

1. **Download** and install WEKA from <http://www.cs.waikato.ac.nz/ml/WEKA/> to the C:\Program Files folder.
2. **Find** and **right click** the “My Computer” or “Computer” icon.
3. **Select** “Properties,” **click** “advanced system settings,” and then **click** “Environmental variable.”
4. Under system variables, **click** “New.”
5. **Type** classpath as your variable name, and C:\Program Files\WEKA-3-6\WEKA-src.jar as your variable value.
6. **Click** “OK.”

WEKA is now installed with all the necessary features that our project requires. Important note: if you did not save WEKA to your Program Files, you cannot use C:\Program Files\WEKA-3-6\WEKA-src.jar as your class variable. Instead, you would replace “Program Files” with the path to the folder where WEKA is downloaded.

Downloading and Optimizing Apache Maven

1. **Download** the zip folder for Apache Maven 3.1.1 from <http://maven.apache.org/download.cgi>.
2. **Extract** the contents of the folder.
3. **Find** and **right click** the “My computer” or “computer” icon.
4. **Select** “Properties,” **click** “advanced system settings,” and then **click** “Environmental variable.”
5. Under system variables, **click** “New.”
6. **Type** “Maven” as your variable name, and the location of your extracted apache maven folder as your variable value.
7. **Click** “OK.”
8. Under system variables, **find** “classpath” and **click** “edit.”
9. Add Maven to your classpath by **typing** Maven into your variable name. Separate it from the preceding item in your classpath with a semicolon.
10. Click “OK.”

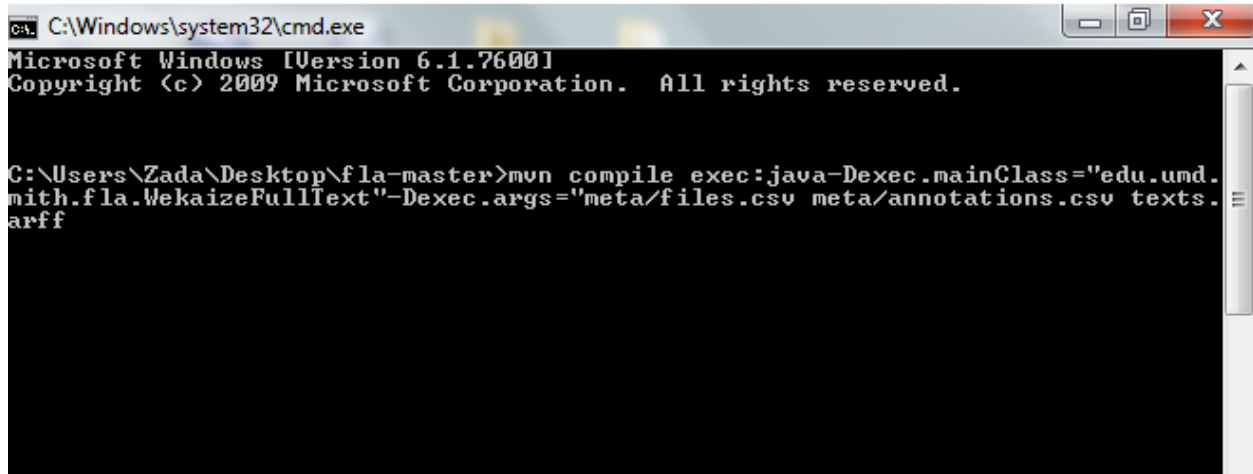
Apache Maven can now be used to generate the ARFF file.

Generating an ARFF file

1. **Visit** <https://github.com/umdmith/fla> and **click** “Download Zip” to download the files to your computer.
2. **Extract** the files to the computer.

3. **Open** the command prompt on the computer and enter the following code while pointing to the fla-master folder:

```
mvn compile exec:java-Dexec.mainClass="edu.umd.mith fla.WekaizeFullText"-  
Dexec.args="meta/files.csv meta/annotations.csv texts.arff"
```



The image shows a screenshot of a Windows command prompt window. The title bar reads "cmd. C:\Windows\system32\cmd.exe". The window content displays the following text:

```
Microsoft Windows [Version 6.1.7600]  
Copyright (c) 2009 Microsoft Corporation. All rights reserved.  
  
C:\Users\Zada\Desktop\fla-master>mvn compile exec:java-Dexec.mainClass="edu.umd.  
mith fla.WekaizeFullText"-Dexec.args="meta/files.csv meta/annotations.csv texts.  
arff
```

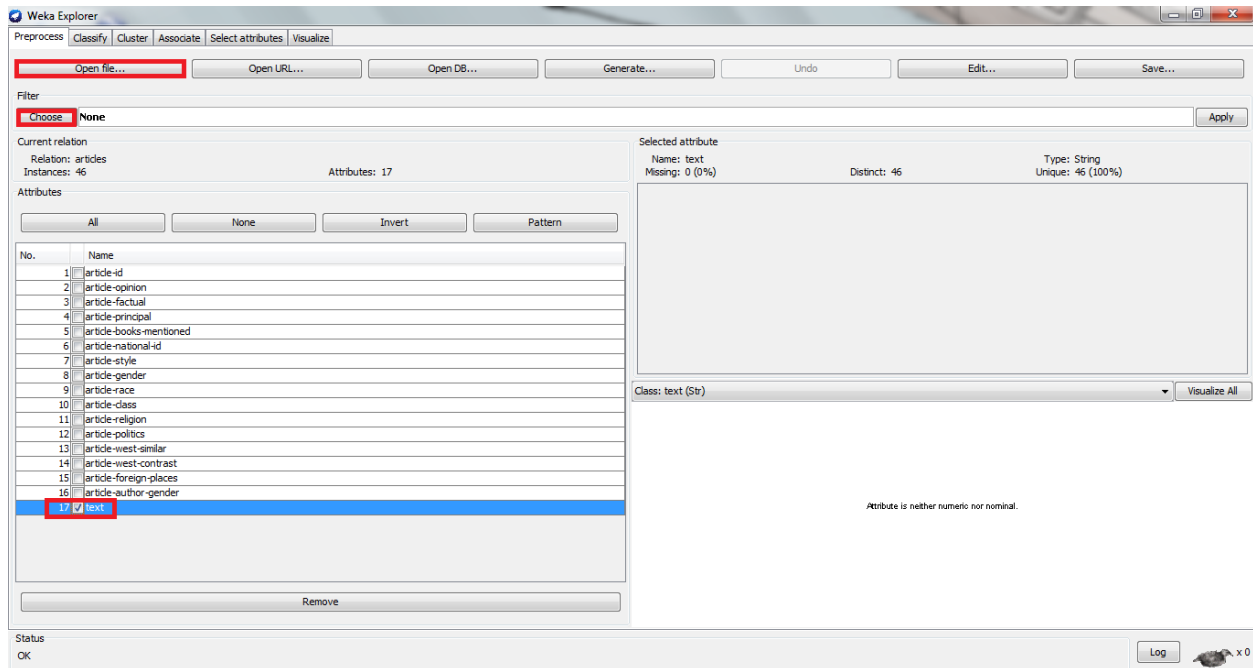
4. **Press** the “Enter” key while in the command prompt menu.

An ARFF file will appear in the fla-master folder. This ARFF file contains the answers to the annotation questions as well as the corresponding texts. The file can be opened using WEKA. We would like to extend our deepest gratitude to Travis Brown of the Maryland Institute of Technology in the Humanities for his work in creating this Github page.

Appendix G: Preprocessing the Data and Using Machine Learning Algorithms

Preprocessing and Filtering the Data

1. **Open WEKA** and **select** the “Explorer” application.
2. **Click** “Open file” and **select** the ARFF file of your choice.
3. **Select** “text” in the Attribute panel.
4. Under the filter section, **select** “Choose,” then “Filters,” then “Unsupervised,” then “Attribute,” then “String-to-word-vector.”



5. **Click** on the “String-to-word-vector” box under the filter category to access the generic object editor. From the generic object editor of the string-to-word-vector algorithm, you can edit the settings for word frequency, stemmers, tokenizers, stop list, and words- to-keep.

To edit word frequencies

6. In the “minTermFreq” box, **type** the desired minimum word frequency. We set the value at 1.

To edit stemmers

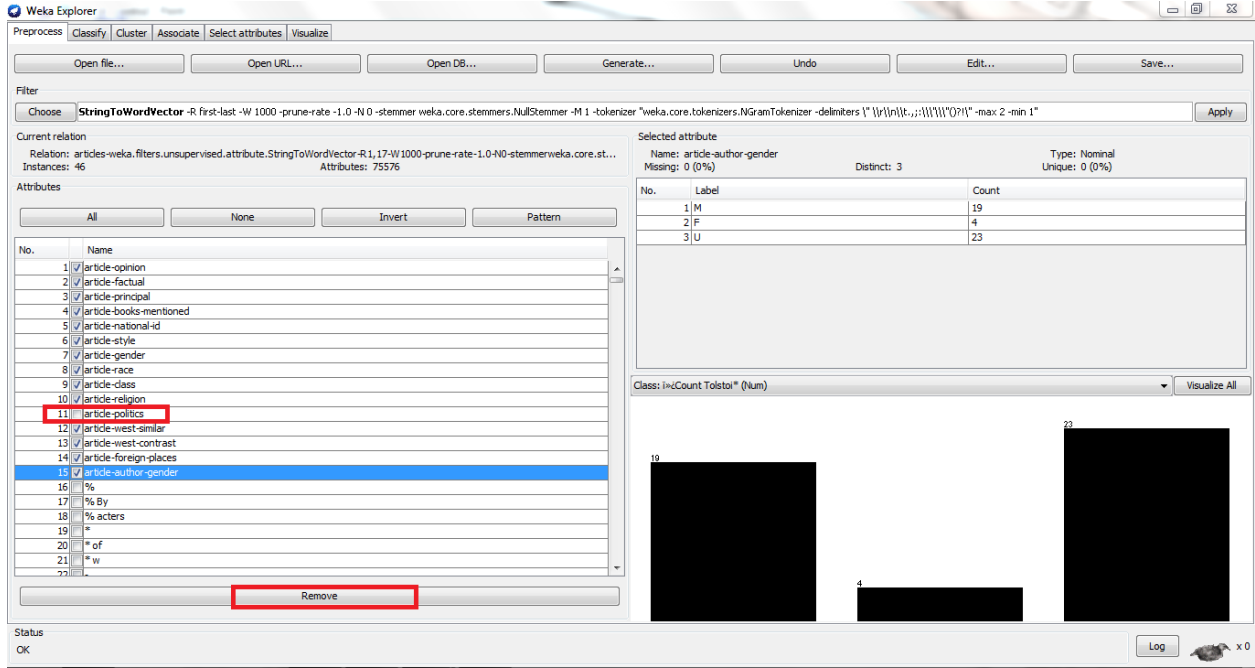
7. In the “stemmer” box, **click** “Choose” and **select** “LovinsStemmer” or “NullStemmer.” Both were used in our experiment.

To edit tokenizers

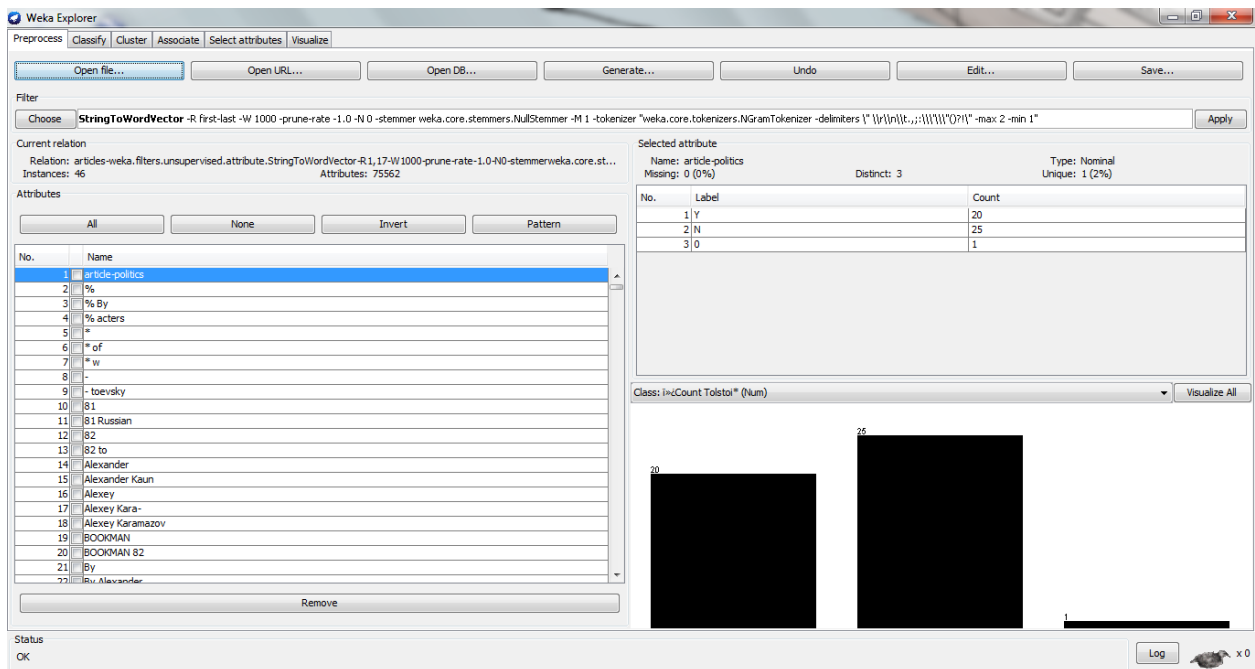
8. To solely use unigrams, **keep** the default “WordTokenizer” option in the tokenizer box. This is the configuration we used in our experiment.

17. In the Attribute panel, **select** all the attributes except the words derived from the text attribute and the one you would like to test. For example, if you would like to classify “article-politics,” you would not select it or any word vectors.

18. **Click** “Remove” below the Attribute panel.

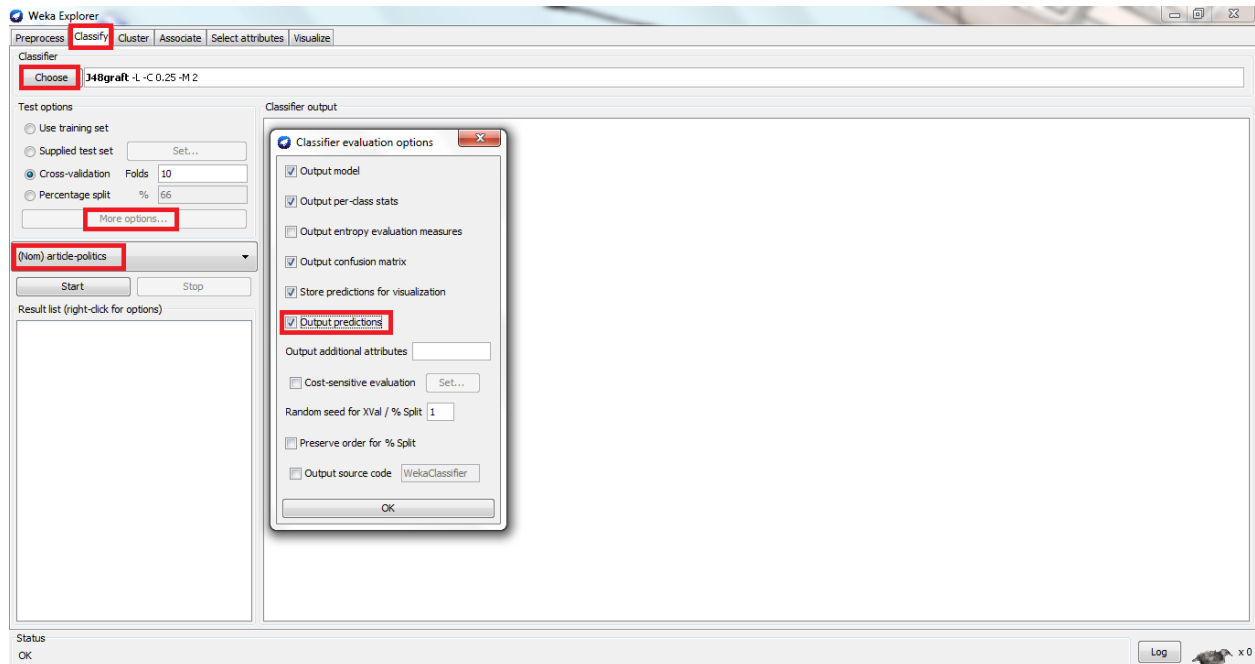


In the Attribute panel, you are left with the attribute you want to classify and all words derived from the text attribute. Now, your data is properly filtered and ready for testing.

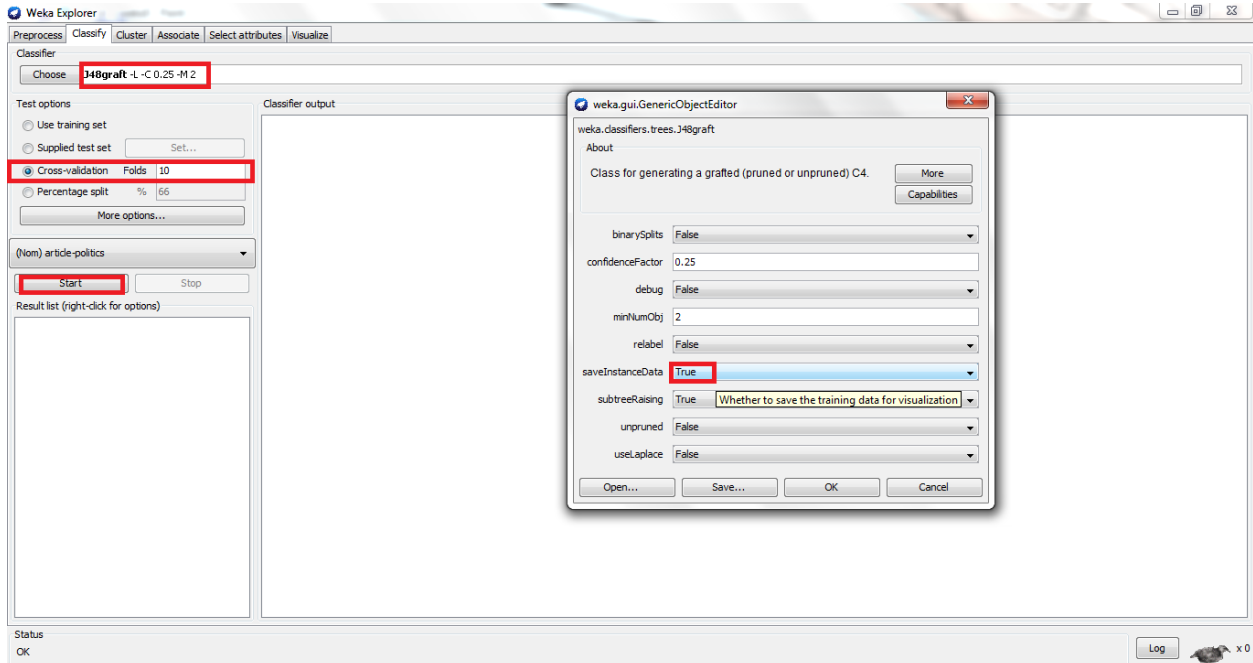


Testing the Training Set

1. **Click** the “Classify tab” at the top of WEKA’s explorer interface.
2. **Select** the attribute you wish to test from the drop-down menu above the “Start” button. You will find the attribute as the first one on the list.
3. **Click** “Choose” from the classifier panel and **select** the classifying algorithm of your choice. For example, J48 decision trees can be found under the “trees” folder and the ZeroR classifier can be found under the “rules” folder.
4. **Click** “More options” in the test options panel and **check** output predictions.



5. **Click** the “J48” box in the classifier panel and **change** “save-instance-data” from false to true. Steps 4 and 5 will allow WEKA to track classifications for each individual instance.
6. In the Test options panel, **select** your test option of choice. A cross validation with 10 folds is preferred.
7. **Click** “Start” above the result list to run the experiment.



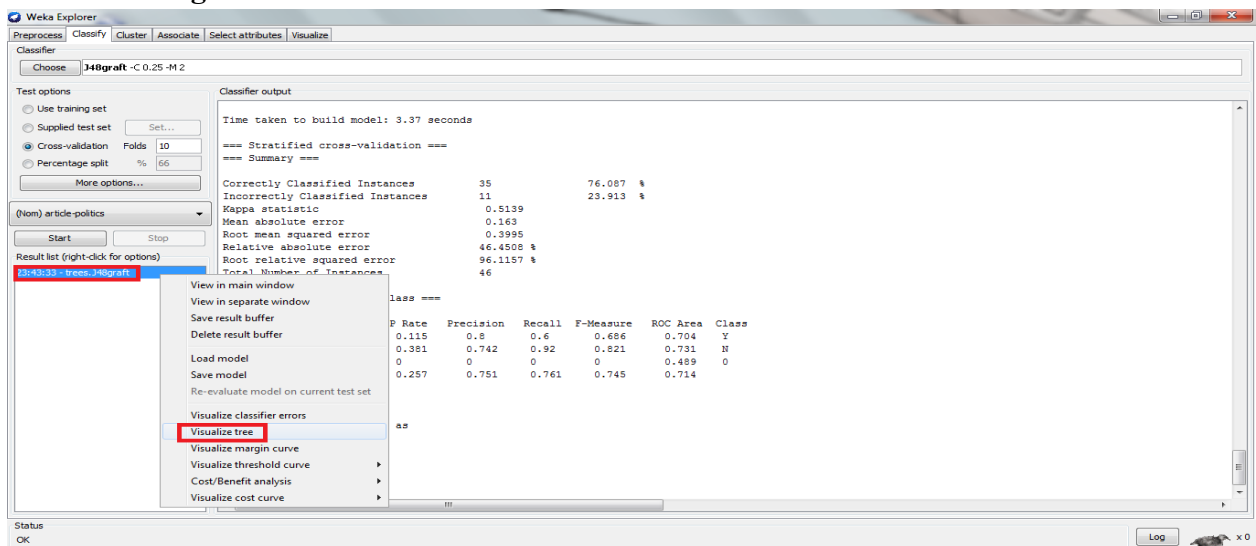
Depending on the size of the dataset, WEKA may take a few minutes to run the experiment. The bird icon at the bottom right will stop moving once the experiment is complete.

Analyzing the Results

The decision tree presented in the classifier output panel is in a text format and thus difficult to navigate. Using a visual model is easier.

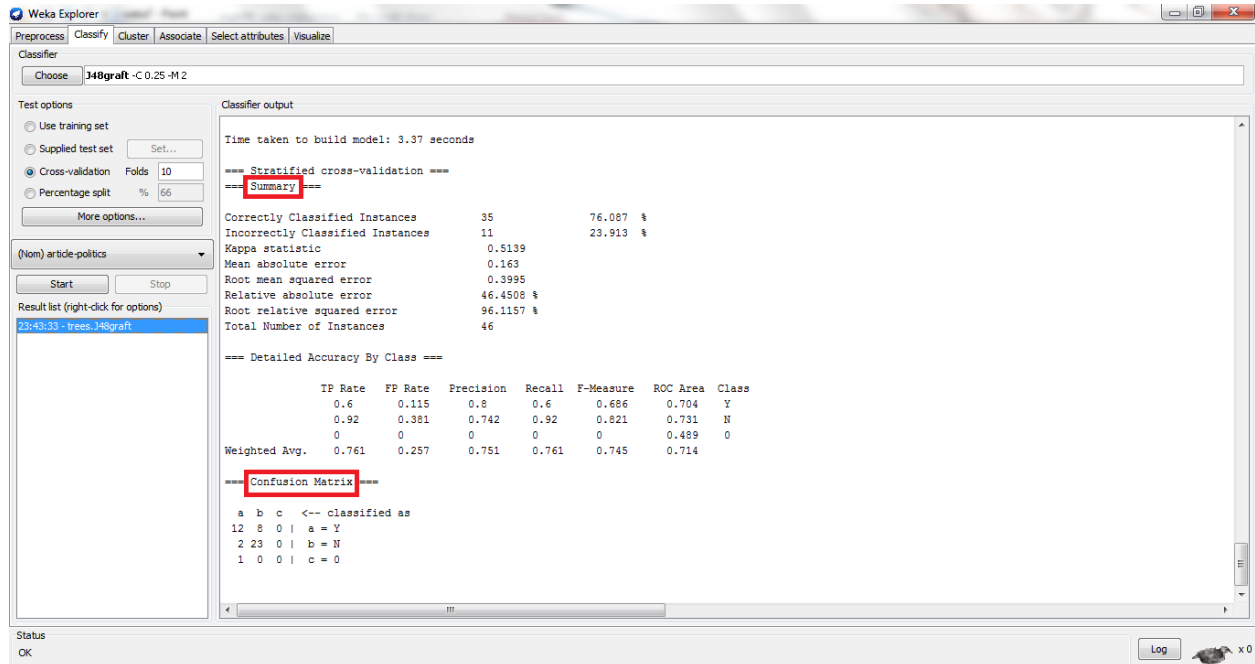
To visualize a decision tree

1. **Right-click** “trees.j48” in the results list panel. **Select** “visualize tree.” A new window opens.
2. **Maximize** the new window.
3. **Right click** the white space and **select** “Auto scale.” Then **right click** again and select “Center to top.” The decision tree is now properly scaled.
4. **Click and drag** the mouse to follow the decision tree.



To access quantitative information

5. In the classifier output panel, **scroll down** until you reach the “Summary” section. You will find important statistical data such as percent of correctly classified instances (i.e. the accuracy rate) and ROC area. This information will reveal the accuracy of the decision tree.
6. Below the summary section you will find the confusion matrix to see how many instances were incorrectly classified.



7. **Scroll up** until you reach the “Predictions on test data” section. This section reveals which instances were incorrectly classified by WEKA.

To save the model for future use and to reexamine old models

8. **Right-click** “trees.j48” in the results list panel and **select** “save model.”
9. **Right-click** “trees.j48” in the results list panel and **select** “load model” to access a previously saved model.

Appendix H: Partial Rotated Component Matrix for Factor Analysis Data (Factors as Columns)

		1	2	3	4	5	6	7	8	9
164										
165										
166										
167	topic-00	.090	-.234	.283	-.148	-.198	.139	.306	.066	.013
168	topic-01	.025	-.038	.107	-.041	-.009	.077	.063	.025	.087
169	topic-02	-.162	.665	.054	-.012	-.026	-.101	-.122	.078	.070
170	topic-03	.036	-.073	.013	.750	.088	.000	.029	.045	.072
171	topic-04	.000	-.016	.039	-.049	-.031	.033	.044	.008	-.003
172	topic-05	.052	-.073	.089	-.043	-.018	.071	-.082	-.011	.039
173	topic-06	.056	-.048	-.773	-.019	-.036	.085	-.017	.000	.088
174	topic-07	.014	-.005	.058	-.017	-.071	.060	-.821	.031	-.068
175	topic-08	.042	-.055	.122	.109	-.158	-.042	.066	-.064	-.098
176	topic-09	.092	-.033	.046	-.081	.005	-.839	.038	.039	.122
177	topic-10	-.126	-.177	-.082	-.060	-.231	-.411	.217	.016	-.127
178	topic-11	-.007	-.023	.021	-.031	.001	.013	.005	.032	.006
179	topic-12	.043	-.050	.060	-.067	-.019	.075	.048	.043	.082
180	topic-13	.001	-.016	-.012	-.035	-.025	.019	.011	.028	-.046
181	topic-14	.079	-.044	.154	-.087	-.043	.085	-.047	.026	-.835
182	topic-15	-.040	-.020	.014	-.070	-.014	.035	.016	-.918	.036
183	topic-16	.090	.218	.117	.619	-.006	.120	.013	.069	.002
184	topic-17	-.383	-.060	.070	.123	.299	.034	.147	-.163	.157
185	topic-18	.017	-.029	.023	-.026	.857	-.025	.081	-.010	.033
186	topic-19	-.048	-.035	-.057	-.049	-.034	.005	.052	.070	-.054
187	topic-20	.140	.561	-.061	-.013	.069	.135	-.039	.103	-.393
188	topic-21	.041	-.011	-.031	-.058	.019	.034	-.014	.029	.005
189	topic-22	.018	-.010	-.051	-.059	.113	.076	-.095	.059	.107
190	topic-23	.038	-.124	-.636	-.092	-.173	-.192	.242	.018	.134
191	topic-24	.112	.754	.055	.039	-.042	.086	.115	-.068	.049
192	topic-25	-.636	-.072	.016	-.024	.062	-.024	.060	-.385	-.019
193	topic-26	.183	-.232	.303	-.139	-.229	-.135	-.073	.166	.179
194	topic-27	-.818	.003	.082	-.111	-.014	.094	.006	.145	.103
195	topic-28	.166	-.184	-.069	.352	-.122	.230	-.009	.105	.240
196	topic-29	-.011	-.011	.056	-.011	-.025	.052	.025	.020	-.013
197	topic-30	.253	-.123	.339	-.137	-.051	.024	-.431	-.024	.276
198	topic-31	.182	.067	-.398	-.258	.159	.216	-.046	.094	-.037
199	topic-32	-.001	-.034	.046	-.019	.013	.034	.025	.016	.009
200	topic-33	-.001	-.064	-.060	.138	-.061	.019	.043	.006	-.091
201	topic-34	-.166	-.017	.115	.257	.534	.103	-.054	.055	-.040
202	topic-35	-.029	-.027	.025	-.038	-.016	.010	.029	.040	-.012
203	topic-36	.041	.007	-.032	-.004	-.018	-.027	-.036	.045	.005
204	topic-37	.302	-.036	.002	-.325	-.195	-.233	.106	.187	-.084
205	topic-38	.372	-.047	.069	-.215	-.056	.299	.245	.244	.224
206	topic-39	-.010	-.061	.030	.011	-.035	-.059	.027	.008	-.004
207	Extraction Method: Principal Component Analysis.									
207	Rotation Method: Varimax with Kaiser Normalization.									
208	a. Rotation converged in 19 iterations.									

Appendix I: Results of All Four Decision Tree Configurations for Each Annotation Topic

<i>Principal Author Subject of Debate</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	73.0290	73.0290	73.0290	73.0290
J48 Accuracy (%)	65.5602	66.39	59.751	74.2739
J48 ROC Area	0.541	0.564	0.494	0.651

<i>Books Mentioned</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	79.6680	79.6680	79.6680	79.6680
J48 Accuracy (%)	70.1245	70.1245	66.3900	68.4647
J48 ROC Area	0.598	0.547	0.555	0.497

<i>Nationality</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	64.7303	64.7303	64.7303	64.7303
J48 Accuracy (%)	71.3693	70.1245	68.0498	68.8797
J48 ROC Area	0.698	0.662	0.685	0.653

<i>Style</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	49.7925	49.7925	49.7925	49.7925
J48 Accuracy (%)	59.3361	65.9751	62.2407	64.3154
J48 ROC Area	0.585	0.659	0.605	0.676

<i>Gender</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	65.1452	65.1452	65.1452	65.1452
J48 Accuracy (%)	59.3361	65.1452	67.2199	63.9004
J48 ROC Area	0.530	0.604	0.648	0.570

<i>Race</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	70.5394	70.5394	70.5394	70.5394
J48 Accuracy (%)	64.7303	74.2739	73.029	68.8797
J48 ROC Area	0.572	0.690	0.640	0.608

<i>Class</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	55.6017	55.6017	55.6017	55.6017
J48 Accuracy (%)	62.6556	67.6349	60.1660	59.3361
J48 ROC Area	0.623	0.662	0.564	0.565

<i>Religion</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	56.4315	56.4315	56.4315	56.4315
J48 Accuracy (%)	60.9959	61.8257	67.6349	62.2407
J48 ROC Area	0.603	0.611	0.638	0.620

<i>Politics</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	65.5602	65.5602	65.5602	65.5602
J48 Accuracy (%)	70.5394	75.1037	75.1037	75.1037
J48 ROC Area	0.643	0.730	0.716	0.741

<i>West Similar</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	66.3900	66.3900	66.3900	66.3900
J48 Accuracy (%)	67.6349	58.0620	61.8257	71.7842
J48 ROC Area	0.611	0.505	0.575	0.639

<i>West Contrast</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	64.7303	64.7303	64.7303	64.7303
J48 Accuracy (%)	71.3693	61.8257	70.9544	64.7303
J48 ROC Area	0.654	0.577	0.696	0.597

<i>Foreign Place Names</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	78.8382	78.8382	78.8382	78.8382
J48 Accuracy (%)	67.2199	73.8589	71.7842	69.2946
J48 ROC Area	0.557	0.606	0.577	0.600

<i>Author's gender</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	85.9375	85.9375	85.9375	85.9375
J48 Accuracy (%)	79.6875	76.5625	71.8750	76.5625
J48 ROC Area	0.509	0.595	0.482	0.403

<i>Opinion: Positive vs. Negative vs. Neutral/Mixed Opinion</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	53.5270	53.5270	53.5270	53.5270
J48 Accuracy (%)	48.9627	57.6763	46.8880	55.1867
J48 ROC Area	0.507	0.591	0.483	0.543

<i>Opinion: Positive vs. Negative</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	91.9643	91.9643	91.9643	91.9643
J48 Accuracy (%)	87.5000	89.2857	83.9286	87.5000
J48 ROC Area	0.564	0.502	0.451	0.348

<i>Opinion: Positive vs. Non Positive (i.e. Negative and Neutral/Mixed Opinion)</i>	Configuration 1	Configuration 2	Configuration 3	Configuration 4
ZeroR Accuracy (%)	57.2614	57.2614	57.2614	57.2614
J48 Accuracy (%)	50.6224	60.9959	51.0373	52.2822
J48 ROC Area	0.468	0.588	0.518	0.498

Appendix J: Topic Modeling Experiment 1 Tables

Factors Only Predictive Accuracy

Step	out1920_1922	<=1919	527	84	86.3
17		>=1923	160	265	62.4
	Overall Percentage				76.4

Factors Only SBLR Model

		B	S.E.	Wald	df	Sig.	Exp(B)
Step	FAC2_1	-.316	.088	13.041	1	.000	.729
17 ^a	FAC3_1	.910	.108	71.265	1	.000	2.484
	FAC4_1	-.189	.087	4.730	1	.030	.827
	FAC5_1	-.216	.077	7.939	1	.005	.806
	FAC6_1	-.201	.073	7.626	1	.006	.818
	FAC7_1	-.182	.082	4.962	1	.026	.834
	FAC8_1	.217	.078	7.824	1	.005	1.242
	FAC9_1	.296	.086	11.927	1	.001	1.345
	FAC10_1	-.518	.085	36.967	1	.000	.596
	FAC11_1	.276	.075	13.565	1	.000	1.318
	FAC12_1	-.677	.100	45.396	1	.000	.508
	FAC13_1	-.467	.093	25.427	1	.000	.627
	FAC14_1	-.148	.070	4.496	1	.034	.863
	FAC15_1	.141	.065	4.655	1	.031	1.151
	FAC18_1	.678	.123	30.420	1	.000	1.970
	FAC19_1	-.448	.105	18.220	1	.000	.639
	FAC21_1	.510	.191	7.111	1	.008	1.666
	Constant	-.391	.083	22.183	1	.000	.676

Topics Only Predictive Accuracy

Step	out1920_1922	<=1919	543	68	88.9
17		>=1923	145	280	65.9
	Overall Percentage				79.4

Topics Only SBLR Model

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 17 ^q	topic02	-9.791	3.859	6.436	1	.011	.000
	topic03	4.066	1.683	5.836	1	.016	58.325
	topic05	10.151	1.687	36.206	1	.000	25607.345
	topic06	-	2.927	12.277	1	.000	.000
		10.256					
	topic07	4.772	1.721	7.693	1	.006	118.167
	topic08	13.372	2.801	22.797	1	.000	641900.416
	topic10	-6.613	1.582	17.470	1	.000	.001
	topic11	17.931	9.372	3.660	1	.056	61262658.900
	topic12	8.465	3.101	7.452	1	.006	4745.608
	topic13	6.352	3.023	4.415	1	.036	573.877
	topic14	28.392	3.491	66.147	1	.000	2139516335559.500
	topic20	-8.171	1.625	25.297	1	.000	.000
	topic26	-4.126	1.407	8.604	1	.003	.016
	topic28	-4.350	2.193	3.936	1	.047	.013
	topic36	5.555	2.060	7.270	1	.007	258.508
	topic37	8.646	1.665	26.975	1	.000	5689.288
	topic39	20.372	6.359	10.262	1	.001	703489788.902
	Constant	-.662	.205	10.470	1	.001	.516

Bibliography (MLA)

- Anschel, Eugene. *The American Image of Russia, 1775-1917*. New York: Ungar, 1974. Print.
- Anta, A., et al. "Sentiment Analysis and Topic Detection of Spanish Tweets: a Comparative Study of NLP Techniques." *La Revista de Procesamiento de Lenguaje* 50 (2013): 45-52. Web. 21 Feb. 2014.
- Astroff, Roberta J. "Revitalizing a Foreign Literature Collection." *Collection Building* 20.1 (2001): 11-18. *ProQuest*. Web. 8 Sept. 2011.
- Aubry, Timothy. "Afghanistan Meets the Amazon: Reading the Kite Runner in America." *PMLA: Publications of the Modern Language Association of America* 124.1 (2009): 25-43. *EBSCO*. Web. 10 Sept. 2011.
- Bai, Jushan, and Serena Ng. "Large Dimensional Factor Analysis." *Foundations and Trends in Econometrics*. 3.2 (2008): 89-163. Web. 14 Mar. 2014.
- Bekkerman, F., and J. Allan. "Using Bigrams in Text Categorization." *Center of Intelligent Information Retrieval* 408 (2003): 1-10. *Cite Seer X*. Web. 23 Nov. 2012.
- Berson, Alex, Stephen Smith, and Kurt Thearling. *Building Data Mining Applications for CRM*. New York: McGraw Hill, (1999): n. pag. Print.
- Blei, David M. "Topic Modeling and Digital Humanities." *Journal of Digital Humanities* 2.1 (2012): n. pag. Web. 1 Mar. 2014.
- Blei, David M., and John D. Lafferty. "Topic models." Department of Computer Science, Princeton University (2009): 1-24. Web. 22 Dec. 2013.
- Block, Sharon. "Doing More with Digitization." *Common-place: The Interactive Journal of Early American Life* 6.2 (2006): n. pag. Web. 1 Mar. 2014.
- Bouckaert, Remco R., et al. "WEKA--Experiences With A Java Open-Source Project." *Journal Of Machine Learning Research* 11.9 (2010): 2533-41. *Academic Search Premier*. Web. 23 Nov. 2012.
- Boyer, Mark A. "Issue Definition and Two-Level Negotiations: An Application to the American Foreign Policy Process." *Diplomacy & Statecraft* 11.2 (2000): 185-212. *America: History and Life with Full Text*. Web. 27 Nov. 2011.
- Bradley, A. "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms." *Pattern Recognition* 30.7 (1997): 1145-59. *Cite Seer X*. Web. 21 Feb. 2014.
- Brands, Hal. "Economic Development and the Contours of U.S. Foreign Policy: The Nixon Administration's Approach to Latin America, 1969-1974." *Peace & Change* 33.2 (2008): 243-73. *Academic Search Premier*. Web. 27 Nov. 2011.
- Brett, Megan R. "Topic Modeling: A Basic Introduction." *Journal of Digital Humanities* 2.1 (2012) : n. pag. Web. 1 Mar. 2014.
- Brown, Archie, Michael C. Kaser, and Gerald S. Smith. *The Cambridge Encyclopedia of Russia and the Former Soviet Union*. Cambridge: Cambridge University Press, 1994. Print.
- "China's Loss of Sovereignty in Manchuria 1895 - 1914." *Today in History, Birthdays & Historical Events*. N.p., n.d. Web. 18 Mar. 2014.
- Cohen, Daniel. "By the Book: Assessing the Place of Textbooks in U.S. Survey Courses." *The Journal of American History* 92.2 (2005): 1405-15. Print.
- Cohn, Deborah. "A Tale of Two Translation Programs." *Latin American Research Review* 41.2 (2006): 139-64. *Academic Search Premier*. Web. 15 Nov. 2011.
- Corse, Sarah M. *Nationalism and Literature: The Politics of Culture in Canada and the United States*. Cambridge: Cambridge University Press, 1997. Print.

- . "Nations and Novels: Cultural Politics and Literary Use." *Social Forces* 73.4 (1995): 1279-308. *JSTOR*. Web. 8 Sept. 2011.
- Croce, Danilo. "Decision Tree Algorithm: WEKA Tutorial." Web Lecture. 21 Feb. 2014.
- Davis, Donald E, and Eugene P. Trani. *Distorted Mirrors: Americans and Their Relations with Russia and China in the Twentieth Century*. Columbia: University of Missouri Press, 2009. Print.
- EBSCOhost. "Readers' Guide Retrospective: 1890-1982, Complete Coverage of Important History-making Events." *H. W. Wilson Databases* (2014). Web. 19 Mar. 2014.
- Foglesong, David S. *The American Mission and the "Evil Empire": The Crusade for a "free Russia" Since 1881*. New York: Cambridge University Press, 2007. Print.
- The Foreign Commerce and Navigation of the United States for the Year Ending [1894, 1897, 1904, 1965]*. Washington, D.C.: G.P.O, 1894—. Print.
- Gaddis, John L. *Russia, the Soviet Union, and the United States: An Interpretive History*. New York: Wiley, 1978. Print.
- Gilens, Martin. "Political Ignorance and Collective Policy Preferences." *American Political Science Review*. 95.2 (2001): 379-96. Web. 29 Nov. 2011.
- Graham, Shawn, and Ian Milligan. "Review of MALLETT, Produced by Andrew Kachites McCallum." *Journal of Digital Humanities* 2.1 (2012) : n. pag. Web. 1 Mar. 2014.
- Grayson, Benson L. *The American Image of Russia, 1917-1977*. New York: Frederick Ungar Pub, 1978. Print.
- Graziano, Anthony M. and Michael L. Raulin. *Research Methods: A Process of Inquiry*. 6th ed. Boston: Pearson Education, Inc., 2007. Print.
- Griswold, Wendy. "The Fabrication of Meaning: Literary Interpretation in the United States, Great Britain, and the West Indies." *American Journal of Sociology* 92.5 (1987): 1077-115. *JSTOR*. Web. 13 Sept. 2011.
- Hall, David, Daniel Jurafsky, and Christopher D. Manning. "Studying the History of Ideas Using Topic Models." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.
- Haslam, Paul Alexander. "The Evolution of the Foreign Direct Investment Regime in the Americas." *Third World Quarterly* 31.7 (2010): 1181-203. *Academic Search Premier*. Web. 27 Nov. 2011.
- "Humanistic Computing." *Proceedings of the IEEE* 86.11 (1998): 2123-51. Web. 15 Nov. 2012.
- Lakshmi, R., and N. Radha. "Supervised Learning Approach for Spam Classification Analysis using Data Mining Tools." *International Journal on Computer Science and Engineering* (2010): 2783-9. *Cite Seer X*. Web. 23 Nov. 2012.
- Larsen, Neil. "The 'Boom' Novel and the Cold War in Latin America." *Modern Fiction Studies* 38.3 (1992): 771-84. *Academic Search Premier*. Web. 16 Nov. 2011.
- Lee, K., et al. "Twitter Trending Topic Classification." *IEEE Data Mining Workshops at ICDM* (2011): 252-58. *ACM Digital Library*. Web. 21 Feb. 2014.
- Levering, Ralph B. *American Opinion and the Russian Alliance, 1939-1945*. Chapel Hill: University of North Carolina Press, 1976. Print.
- Levinson, Brett. *The Ends of Literature: The Latin American 'Boom' in the Neoliberal Marketplace*. Stanford, CA: Stanford UP, 2001. *MLA International Bibliography*. Web. 27 Nov. 2011.
- Li, V. "Misgivings of a Tongue-Tied Nation." *Editorial Research Reports* 2 (1990): n. pag. Web. *CQ Researcher*. 13 Sept. 2011.

- Majnik, M., and Z. Bosnic. "ROC Analysis of Classifiers in Machine Learning: a Survey." *Technical Report MM-1 University of Ljubljana* (2011): 1-34. Web. 21 Feb. 2014.
- Mallios, Peter Lancelot. *Our Conrad: Constituting American Modernity*. Stanford: Stanford UP, 2010. *Google Books*. Web. 15 Sept. 2011.
- Mann, Steve. "Humanistic Intelligence: 'WearComp' as a New Framework and Application for Intelligent Signal Processing." *Proceedings of the IEEE* 86.11 (1998): 2123-53. Web. 14 Nov. 2013.
- Martin, Gerald. "Boom, Yes; 'New' Novel, No: Further Reflections on the Optical Illusions of the 1960s in Latin America." *Bulletin of Latin American Research* 3.2 (1984): 53-63. Web. 25 Oct. 2011.
- May, Rachel. *The Translator in the Text: on Reading Russian Literature in English*. Evanston: Northwestern UP, 1994. Print.
- Meeks, Elijah, and Scott B. Weingart. "The Digital Humanities Contribution to Topic Modeling." *Journal of Digital Humanities* 2.1 (2012) : n. pag. Web. 1 Mar. 2014.
- Newton, Douglas "World War I." *Berkshire Encyclopedia of World History, Second Edition*. Great Barrington: Berkshire Publishing Group, 2011. Credo Reference. Web. 19 March 2014.
- Nuntiyagul, Atorn, et al. "Keyword Extraction Strategy For Item Banks Text Categorization." *Computational Intelligence* 23.1 (2007): 28-44. *Academic Search Premier*. Web. 21 Feb. 2014.
- Ockerbloom, Mark John, ed. *The Online Books Page: Serials*. Philadelphia: U of Pennsylvania, (1997): n. pag. Web. 29 Nov. 2011.
- Ohmann, Richard. "The Shaping Of A Canon: U.S. Fiction, 1960-1975." *Critical Inquiry* 10.1 (1983): 199-223. *MLA International Bibliography*. Web. 13 Nov. 2011.
- Paynter, Gordon. "Attribute-Relation File Format." *Department of Computer Science at the University of Waikato* (2008): n. pag. Web. 18 Feb. 2014.
- Peattie, Mark R. "Russo-Japanese War." *The Reader's Companion to Military History*. Boston: Houghton Mifflin, 1996. Credo Reference. Web. 18 March 2014.
- Peng, F., and Dale Shuurmans. "Combining Naive Bayes and N-gram Language Models for Text Classification." *Proceedings of the 25th European Conference on Information Retrieval Research* (2003): 335-350. *Cite Seer X*. Web. 21 Feb. 2014.
- Perrie, Maureen, D C. B. Lieven, and Ronald G. Suny. *The Cambridge History of Russia*. Cambridge: Cambridge University Press, 2006. Print.
- "Pogroms." *YivoInstitute.org*. YIVO Institute, 2005. Web. 18 Mar. 2014.
<<http://www.yivoInstitute.org/downloads/Pogroms.pdf>>.
- Rhody, Lisa M. "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2.1 (2012): n. pag. Web. 1 Mar. 2014.
- Sable, R., McKeown, K., and Kenneth Church. "NLP Found Helpful (At Least for One Text Categorization Task)." *Conference on Empirical Methods in Natural Language Processing* (2002): 172-9. *ACM Digital Library*. Web. 18 Feb. 2014.
- Sanderson, R., and P. Watry. "Integrating Data and Text Mining Processes for Digital Library Applications." *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (2007): 73-9. *ACM Digital Library*. Web. 23 Nov. 2012.
- Schöch, Christof. "Big? Smart? Clean? Messy? Data in the Humanities." *Journal of Digital Humanities* 2.3 (2013): n. pag. Web. 1 Mar. 2014.

- Shaw, Donald L. "When was Modernism in Spanish-American Fiction?" *Bulletin of Spanish Studies* 79.2-3 (2002): 395-409. *Taylor Francis Online*. Web. 25 Nov. 2011.
- Shlapentokh, Vladimir, Eric Shiraev, and Eero Carroll. *The Soviet Union: Internal and External Perspectives on Soviet Society*. New York, N.Y: Palgrave Macmillan, 2008. Print.
- Simpson, John, et al. "Text Mining Tools in the Humanities: An Analysis Framework." *Journal of Digital Humanities* 2.3 (2013): n. pag. Web. 1 Mar. 2014.
- Singh, Munindar P., ed. *The Practical Handbook of Internet Computing*. Boca Raton: Chapman & Hall, 2005. *Google Books*. Web. 18 Feb. 2014.
- Steyvers, Mark, and Tom Griffiths. "Probabilistic topic models." *Handbook of Latent Semantic Analysis* 427.7 (2007): 424-440.
- Strayer, Robert W. "Revolution—Russia." *Berkshire Encyclopedia of World History, Second Edition*. Great Barrington: Berkshire Publishing Group, 2011. Credo Reference. Web. 18 March 2014.
- Stubbs, Michael. "Conrad in the Computer: Examples of Quantitative Stylistic Methods." *Language and Literature: Journal of the Poetics and Linguistics Association* 14.1 (2005): 5-24. *EBSCO*. Web. 10 Sept. 2011.
- Taddy, Matthew A. "On Estimation and Selection for Topic Models." *Proceedings of AISTATS* (2012): 1184-93. Web. 22 Dec. 2013.
- Tompkins, Jane. *Sensational Designs: the Cultural Work of American Fiction, 1790-1860*. New York: Oxford University Press, 1986. Print.
- Travis, Rick. "Problems, Politics, and Policy Streams: A Reconsideration US Foreign Aid Behavior toward Africa." *International Studies Quarterly* 54.3 (2010): 797-821. *Academic Search Premier*. Web. 27 Nov. 2011.
- Virginia Polytechnic Institute and State University. "Readers' Guide Retrospective: 1890-1982 from EBSCOhost." *University Libraries* (2014). Web. 19 Mar. 2014.
- Wang, Xuerui, and Andrew McCallum. "Topics over Time: a Non-Markov Continuous-Time Model of Topical Trends." *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2006): 424-33. Web. 22 Dec. 2013.
- Watson, Robert P., and Sean McCluskie. "Human Rights Considerations and U.S. Foreign Policy: The Latin American Experience." *Social Science Journal* 34.2 (1997): 249-57. *Academic Search Premier*. Web. 27 Nov. 2011.
- Weeks, Theodore R. *Across the Revolutionary Divide: Russia and the Ussr, 1861-1945*. Chichester, West Sussex: Wiley-Blackwell, 2011. Print.
- Williams, William A. *American-Russian Relations, 1781-1947*. New York: Rinehart, 1952. Print.
- Witten, I. H., Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann, 2011. Print.
- Yi, Xing, and James Allan. "A Comparative Study of Utilizing Topic Models for Information Retrieval." *Department of Computer Science at the University of Massachusetts at Amherst* (2009): 29-41. Web. 22 Dec. 2013.
- Zabriskie, Edward H. *American-Russian Rivalry in the Far East: A Study in Diplomacy and Power Politics, 1895-1914*. Westport, Conn: Greenwood Press, 1973. Print.
- Zhou, Xiaohua, Xiaodan Zhang, and Xiaohua Hu. "Semantic Smoothing of Document Models for Agglomerative Clustering." *College of Information Science and Technology at Drexel University* (2007): 1-6. Web. 22 Dec. 2013.

"1905 Russian Revolution." *Spartacus Educational*. N.p., n.d. Web. 18 Mar. 2014.
<<http://www.spartacus.schoolnet.co.uk/RUS1905.htm>>.