APPENDIX:

"USE OF FORCE AND RUSSIAN CIVIL-MILITARY RELATIONS: AN AUTOMATED CONTENT ANALYSIS"

Brandon M. Stewart and Yuri M. Zhukov

1 April 2009

CONTENTS

APPENDIX A: SOURCES FOR RUSSIAN DOCUMENTS

RUSSIAN SOURCES USED (EAST-VIEW DATABASE)

Military and Security Periodicals

- *Agentstvo voennykh novostei*
- *Armeiskii sbornik*
- *Boevaia vakhta*
- *Flag Rodiny*
- *Krasnaia zvezda*
- *Krasnyi voin*
- *Morskaia gazeta*
- *Morskoi sbornik*
- *Na boevom postu*
- *Na strazhe Rodiny*
- *Nezavisimoe voennoe obozrenie*
- *Novosti razvedki i kontrrazvedki*
- *Orientir*
- *Rossiiskoe voennoe obozrenie*
- *Shchit i mech*
- *Soldat Otechestva*
- *Soldat Rossii*
- *Strazh Baltiki*
- *Tekhnika i vooruzhenie*
- *Ural'skie voennye vesti*
- *VPK. Voenno-promyshlennyi kur'er*
- *VVS segodnia*
- *Vestnik voennoi informatsii*
- *Voennaia mysl'*
- *Voenno-istoricheskii zhurnal*
- *Voenno-meditsinskii zhurnal*
- *Voennye znaniia*
- *Voennyi diplomat*
- *Voennyi vestnik Iuga Rossii*
- *Voin Rossii*
- *Voprosy bezopasnosti*
- *Zakavkazskie voennye vedomosti*
- *Zarubezhnoe voennoe obozrenie*
- *Zashchita i bezopasnost'*

Government Publications

- *Diplomaticheskii vestnik*
- *Gosudarstvennaia Duma. Dnevnik zasedanii*
- *Gosudarstvennaia Duma. Parlamentskie slushaniia*
- *Gosudarstvennaia Duma. Povestka zasedanii Soveta Dumy*
- *Gosudarstvennaia Duma. Stenogramma zasedanii*
- *Parlamentskaia gazeta*
- *Prezident. Poslaniia*
- *Prezident. Vystupleniia*
- *Rossiiskaia gazeta*
- *Sovet Federatsii. Dnevnik zasedanii*
- *Sovet Federatsii. Obrashcheniia*
- *Sovet Federatsii. Stenogrammy zasedanii*
- *Sovet Federatsii. Zaiavleniia*

Central Newspapers

- *Argumenty i fakty*
- *Ekspert*
- *Ezhenedel'nyi zhurnal*
- *Finansovye Izvestiia*
- *Gazeta*
- *ITAR-TASS Daily*
- *InterFax-Vremia*
- *Itogi*
- *Izvestiia*
- *Kommersant. Daily*
- *Kommersant. Vlast'*
- *Komsomol'skaia pravda*
- *Konservator*
- *Krasnaia zvezda*
- *Kul'tura*

Central Newspapers (cont'd)

- *Literaturnaia gazeta*
- *Moskovskaia pravda*
- *Moskovskie novosti*
- *Moskovskii komsomolets*
- *Nasha versiia*
- *New Times*
- *Nezavisimaia gazeta*
- *Novaia gazeta*
- *Novoe vremia*
- *Novye Izvestiia*
- *Ogonek*
- *Pravda*
- *Profil'*
- *Rossiia*
- *Rossiiskaia gazeta*
- *Rossiiskie vesti*
- *Russkii Telegraf*
- *Russkii kur'er*
- *Sankt-Peterburgskie vedomosti*
- *Segodnia*
- *Slovo*
- *Sovetskaia Rossiia*
- *Tribuna*
- *Trud*
- *Vecherniaia Moskva*
- *Vedomosti*
- *Vek Versiia*
- *Vremia MN*
- *Vremia novostei*
- *Zavtra*

APPENDIX B: CODEBOOK

Unit of Analysis: Each observation represents a speech, press briefing, published article, interview or other public statement made by a political or military public figure in Russia. The original dataset included 7,920 observations, 5,143 of which remained in the cross-sectional dataset after the extraction of missing values ("NA" or "other topic/classification") for the dependent variables **TOPIC** and **CLASS**. The period of observation is 12 February 1998 to 31 October 2008, which represent the release dates of the oldest and most recent statements in the collection.

This dataset is used for Models 1-3.

DOCUMENT_ID

Unique code assigned to each observation

DATE

Date of public statement's release. Format: YYYYMMDD

NAME

Surname of author of statement

TS.SELECTION

0=Not included in training set

1=Included in training set (random selection of 300 observations)

TS.CLASS

0=Conservative (training set data only)

1=Activist (training set data only)

99=Not applicable/None of the above (training set data only)

NA=Not included in training set

Statements consistent with a *conservative* outlook (TS.CLASS=0) include:

- Expressions of a preference toward the use of military instruments of power only as a last resort:

  Example: "К военной мощи следует прибегать, когда возможности других средств исчерпаны [Military power should be resorted to when the capabilities of other instruments are exhausted]"

- Expressions of a preference toward interagency solutions to security problems:

  Example: "Военная безопасность Российской Федерации должна достигаться всей деятельностью государства - как в военной сфере, так и в политико-дипломатической, экономической, культурной, социальной и других [The military security of the Russian Federation should be attained through government-wide action — in the military sphere, as well as political-diplomatic, economic, cultural, social and others].

- Explicit doubts about the efficacy of military solutions in addressing specific political problems.

  Example: “С террором необходимо бороться адекватными, но конституционными способами [Terror should be fought with adequate, but constitutional means].”

- Support for multilateral solutions to political and security problems.

  Example: “Решение этих вопросов, особенно решение о применении силы, должно проходить исключительно через Совет Безопасности Организации Объединенных Наций [Decisions on such questions, particularly on the use of force, should be made exclusively through the United Nations Security Council].”

- General support for limited ends and means in foreign and defense policy.

  Example: “Требуется определенная умеренность в определении и отстаивании национальных интересов, чтобы жестко отстаивать только действительно жизненно важные из них [A certain restraint is necessary in the definition and defense of national interests, so as to stringently defend only those, which are truly fundamental].”

Statements consistent with an *activist* outlook (TS.CLASS=1) include:

- Explicit support for the use of military power in a specific scenario

  Example: “Выполняя свой воинский долг, вы хорошо понимали, что вы – это, по сути, последняя надежда беззащитных людей [While performing your military duty, you understood well that you, in essence, were the last hope of a defenseless people].”

- Support for military as most effective instrument of power in a given scenario.

  Example: “Мы часто получаем официальные отзывы представителей ООН и ОБСЕ, которые признают наших миротворцев главным гарантом мира и стабильности [We often receive official feedback from UN and OSCE representatives, who admit that our peacekeepers are main guarantors of peace and stability].”

- Interpretation of the use or show of force as an effective means to certain political ends.

  Example: “Применение ВМФ в мирное и военное время ... существенно повысит степень защищенности и реализуемости военно- стратегических и экономических интересов России в Мировом океане, позволит полностью или в значительной мере решить многие социальные, экономические и экологические проблемы приморских регионов Российской Федерации, а также международных отношений и сотрудничества с соседними морскими государствами [The use of the Navy in peacetime and in war ... will significantly increase the level of defensibility and realizability of Russia’s military-strategic and economic interests in the World Ocean, will enable the solution – in full or in significant part — of many social, economic and ecological problems of littoral areas of the Russian Federation, as well as international relations and cooperation with neighboring coastal states].”

- Support for unilateral solutions to interventionist crises.

  Example: “Россия исторически была и останется гарантом безопасности народов Кавказа [Russia historically has been and will remain the guarantor of security for the peoples of the Caucasus].”

- General preference for foreign policy outcomes that revise, rather than reinforce the status quo.

  Example: “Такого положения, такой опасности Россия не может себе позволить [Russia cannot permit such a state of affairs, such a threat to exist].”

CLASS

0=Conservative

1=Activist

NA=Not Applicable

Values assigned by an ensemble of supervised document classifiers (K-Nearest Neighbor, Adaboost.M1, Random Forest and Support Vector Machine), as outlined in Research Design section. TS.CLASS served as training input for document classification.

TOPIC

0=Realpolitik
1=Interventionist
NA=Not Applicable

Values assigned by Expressed Agenda Model as described in Research Design section, according to categories outlined in Table 1.

MILT

0= Author of statement is a Political Elite
1= Author of statement is a Military Elite

Military Elites include active duty and retired officers of field grade and above, who at the time of the statement's publication occupied a position of formal authority or informal influence in the Ministry of Defence or an affiliated institute/agency. Individuals from the following offices are included in the Military Elite category: Office of the Minister of Defence (uniformed personnel only), General Staff, Service/Branch Headquarters, Territorial Commands, high-visibility unit commands (such as peacekeeping contingents), professional military education institutions, military research institutions and the Academy of Military Sciences.

Political Elites include civilians, who at the time of the statement's publication occupied a position of formal authority or informal influence in the Government of Russia or an affiliated institute/agency. Individuals from the following offices are included in the Political Elite category: Executive Office of the President, Government of Russia (Prime Minister's Office), Security Council, Ministry of Foreign Affairs, Federal Security Service, Foreign Intelligence Service, Federation Council, State Duma, and the Council on Foreign and Defense Policy. Only individuals whose policy portfolios include foreign and security affairs – such as senior officials in cabinet ministries or members of the relevant parliamentary committees – were included in this sample.

CABT

1= Author is member of Presidential Administration or Security Council
0= Otherwise

PMNT

1= Author is member of State Duma or Federation Council
0= Otherwise

INFORMAL

0=Formal Authority

1=Informal Influence

In the cases examined here, a value of 0 ("Formal Authority") was assigned for public statements made by individuals who at the time of publications occupied senior positions in the following structures (in alphabetical order):

- Emergency And Civil Defense State Committee
- Federal Assembly of Russia, Federation Council, Committee on CIS Affairs
- Federal Assembly of Russia, Federation Council, Committee on Defence and Security
- Federal Assembly of Russia, Federation Council,
- Federal Assembly of Russia, State Duma, Committee on CIS Affairs
- Federal Assembly of Russia, State Duma, Committee on Defence
- Federal Assembly of Russia, State Duma, Committee on Federal Budget Appropriations for Defense and State Security
- Federal Assembly of Russia, State Duma, Committee on International Affairs
- Federal Assembly of Russia, State Duma, Committee on Security
- Federal Assembly of Russia, State Duma, Committee on Veterans Affairs
- Federal Assembly of Russia, State Duma, Leadership
- Federal Border Service
- Federal Security Service
- Foreign Intelligence Service
- Government of Russia (Prime Minister's Office)
- Kremlin, Presidential Civil Service Directorate
- Kremlin, Presidential Directorate for Interregional Relations and Cultural Contacts with Foreign Countries
- Kremlin, Presidential Executive Office
- Kremlin, Presidential Foreign Policy Directorate
- Ministry of Defence, Accommodation and Amenity Service, Main Accommodation and Amenity Directorate
- Ministry of Defence, Accommodation and Amenity Service,
- Ministry of Defence, Air Force, 13th State Research Institute
- Ministry of Defence, Air Force, 16th Air Army
- Ministry of Defence, Air Force, 37th Air Army (SOF)
- Ministry of Defence, Air Force, 4th Air and AAD Army
- Ministry of Defence, Air Force, 4th Center for Combat and Flight Training
- Ministry of Defence, Air Force, 5th Air and AAD Army
- Ministry of Defence, Air Force, 61st Air Army (Transport)
- Ministry of Defence, Air Force, 6th Air and AAD Army
- Ministry of Defence, Air Force, Aeronautical Service
- Ministry of Defence, Air Force, Central AF Command Point
- Ministry of Defence, Air Force, Far Eastern Group
- Ministry of Defence, Air Force, Headquarters

- Ministry of Defence, Air Force, Military Academy of Aerospace Defence
- Ministry of Defence, Airborne Troops, Headquarters
- Ministry of Defence, General Staff
- Ministry of Defence, General Staff, Directorate of Communications
- Ministry of Defence, General Staff, Directorate of Military Topography
- Ministry of Defence, General Staff, Directorate of Radio-Electronic Warfare
- Ministry of Defence, General Staff, General Intelligence Directorate
- Ministry of Defence, General Staff, General Mobilization Directorate
- Ministry of Defence, General Staff, General Operational Directorate
- Ministry of Defence, General Staff, Military Academy of the General Staff
- Ministry of Defence, Ground Forces Command, Far Eastern Military District
- Ministry of Defence, Ground Forces Command, Headquarters
- Ministry of Defence, Ground Forces Command, Leningrad Military District
- Ministry of Defence, Ground Forces Command, Moscow Military District
- Ministry of Defence, Ground Forces Command, North Caucasus Military District
- Ministry of Defence, Ground Forces Command, Siberian Military District
- Ministry of Defence, Ground Forces Command, Volga-Ural Military District
- Ministry of Defence, Navy, Baltic Fleet
- Ministry of Defence, Navy, Black Sea Fleet
- Ministry of Defence, Navy, Caspian Flotilla
- Ministry of Defence, Navy, Headquarters
- Ministry of Defence, Navy, Northern Fleet
- Ministry of Defence, Navy, Pacific Fleet
- Ministry of Defence, Office of First Deputy Minister of Defence, Main Directorate for Combat Training and Service
- Ministry of Defence, Office of First Deputy Minister of Defence, Military Inspection
- Ministry of Defence, Office of First Deputy Minister of Defence,
- Ministry of Defence, Office of State Secretary, Directorate of State Civil Service
- Ministry of Defence, Office of State Secretary, Executive and Legislative Powers Cooperation Branch
- Ministry of Defence, Office of State Secretary, General Directorate for Morale
- Ministry of Defence, Office of State Secretary, Main Personnel Directorate
- Ministry of Defence, Office of State Secretary,
- Ministry of Defence, Office of the Deputy Minister,
- Ministry of Defence, Office of the Deputy Minister for Armament, Executive Office
- Ministry of Defence, Office of the Deputy Minister for Armament, General Rocket Artillery Directorate
- Ministry of Defence, Office of the Deputy Minister for Armament, General Tank-Automotive Directorate
- Ministry of Defence, Office of the Deputy Minister for Armament,
- Ministry of Defence, Office of the Deputy Minister for Finiancial and Economic Affairs, Civil Disbursements Directorate
- Ministry of Defence, Office of the Deputy Minister for Finiancial and Economic Affairs, General Finance Directorate
- Ministry of Defence, Office of the Deputy Minister for Finiancial and Economic Affairs
- Ministry of Defence, Office of the Deputy Minister for Logistics, Directorate of Logistics

- Ministry of Defence, Office of the Deputy Minister for Logistics, General Medical Directorate
- Ministry of Defence, Office of the Deputy Minister for Logistics, General Medical Directorate
- Ministry of Defence, Office of the Deputy Minister for Logistics, General Medical Directorate
- Ministry of Defence, Office of the Deputy Minister for Logistics
- Ministry of Defence, Office of the Minister of Defence, Administration of MoD Affairs
- Ministry of Defence, Office of the Minister of Defence, Directorate of Information and Public Relations
- Ministry of Defence, Office of the Minister of Defence, Expert Center
- Ministry of Defence, Office of the Minister of Defence, Financial Inspection
- Ministry of Defence, Office of the Minister of Defence, Main Directorate for International Cooperation
- Ministry of Defence, Office of the Minister of Defence, Main Legal Directorate
- Ministry of Defence, Office of the Minister of Defence, Secretariat
- Ministry of Defence, Office of the Minister of Defence
- Ministry of Defence, Railroad Troops, Headquarters
- Ministry of Defence, Space Forces, Headquarters
- Ministry of Defence, Strategic Rocket Forces, Headquarters
- Ministry of Finance
- Ministry of Foreign Affairs
- Ministry of Internal Affairs
- Ministry of Justice
- Office of the Prosecutor General
- Security Council
- State Narcotics Control Service

A value of 1 ("Informal Influence") was assigned for public statements made by individuals who at the time of publications occupied senior positions in the following structures (in alphabetical order):

- Academy of Military Sciences
- Russian Academy of Sciences
- Council on Foreign and Defence Policy

PFSCR

Composite Press Freedom Score (0-100 ordinal scale). Measured annually.

Source: "Freedom of the Press Historical Data," Freedom House,
http://www.freedomhouse.org/template.cfm?page=274;
"Freedom of the Press 2007 Survey," Freedom House,
http://www.freedomhouse.org/template.cfm?page=107&year=2007;
"Freedom of the Press 2008 Survey," Freedom House,
http://www.freedomhouse.org/template.cfm?page=362.

Coding specifications available at:
http://www.freedomhouse.org/uploads/fop08/Methodology2008.pdf

W1, W2, W3

Three vectors of probabilities of observing Class 1 (Conservative), Class 2 (Activist) or Class 3 (Not Applicable) for each document. These probabilities are obtained from the normalized ensemble classifier weights and can be used to simulate the imputation draws for uncertainty analyses.

DISAGREE

Mean level of disagreement between military and political elites in a given month.

Measured as Cartesian distance between mean monthly military and political elite classifications for CLASS and TOPIC:

$\surd[(\mathrm{MIL.CLASSt} - \mathrm{POL.CLASSt})2 + (\mathrm{MIL.TOPICt} - \mathrm{POL.TOPICt})2]$

SC.MILP

Proportion of Security Council officials with professional military background.

SC.MIL/ SC.TOT

where SC.MIL is the number of members of the Russian Security Council with a professional military background (defined as attendance of a professional military education institution) and SC.TOT is the number of seats on the Security Council.

Source: Security Council of the Russian Federation,
http://www.scrf.gov.ru/persons/sections/parent/

APPENDIX C: DOCUMENT ANALYSIS

This appendix includes supplemental replication details. It generally does not repeat information found in the body of the text, but focuses on additional replication parameters. Full replication code and data will be made available upon request.

Document Analysis Summary

Task	Use of Force	Topic Model
Obtain Source Material	~8000 Press Releases and Interviews (see Appendix D)	
Document Preprocessing	Russian Stemming Feature Extraction	
Feature Selection	Latent Semantic Analysis	Latent Direchlet Allocation
Pre-Analysis Human Intervention	300 Document Set	# of Topics Set
Methods	Ensemble Supervised Learning	Unsupervised Expressed Agenda Model
Post-Processing	Distance Analysis	Topic Interpretation Distance Analysis

C.1: TEXT PROCESSING

We made the decision to analyze texts in their native Russian. Machine translation is notoriously unreliable, despite the relatively relaxed assumptions of bag-of-words analysis. In order to provide the best results, words must be consistent in their meaning across texts. This requires that multiple words meaning different things in Russian not be translated to the same word in English. Unfortunately, machine translation is not even up to simple word-for-word translation that ignores lexical ordering. To use an example from the training set, the sentence "применение ВМФ в мирное и военное время ... существенно повысит степень защищенности ... интересов России" should translate to "employment of the Navy in peacetime and wartime ... would substantially increase the level of *defensibility* ... of Russia's interests." Meanwhile, machine translation (BabelFish) translates the sentence as "the application of the Navies in the peaceful and military time... will significantly increase the *vulnerability* of the ... interests of Russia." Thus, the machine-translated text would receive a coding of "Conservative," since it would appear to argue that force employment makes Russia *less* safe, while the original Russian clearly argues the opposite.

Text processing was performed using a custom Python script written under Python 2.7 to handle multiple encoding formats of the Cyrillic text. The python script was built using the BeautifulSoup HTML parsing library and the Snowball Project Russian stemmer. It creates a *tf*idf* matrix and a count matrix, which include only unigrams which appear in more than 1% of the text and less than

99% of the texts. A list of stop-words and stop-stems were removed from the analysis. The stop-words were generated from a list provided with the snowball stemmer. The stop-stems were generated empirically by the authors in the process of refining the Expressed Agenda model.

Stop-words:

и,в,во,не,что,он,на,я,с,со,как,а,то,все,она,так,его,но,да,ты,к,у,же,вы,за,бы,по,только,ее,мне,было,вот,от,меня,еще,нет,о,из,ему,теперь,когда,даже,ну,вдруг,ли,если,уже,или,ни,быть,был,него,до,вас,нибудь,опять,уж,вам,сказал,ведь,там,потом,себя,ничего,ей,может,они,тут,где,есть,надо,ней,для,мы,тебя,их,чем,была,сам,чтоб,без,будто,человек,чего,раз,тоже,себе,под,жизнь,будет,ж,тогда,кто,этот,говорил,того,потому,этого,какой,совсем,ним,здесь,этом,один,почти,мой,тем,чтобы,нее,кажется,сейчас,были,куда,зачем,сказать,всех,никогда,сегодня,можно,при,наконец,два,об,другой,хоть,после,над,больше,тот,через,эти,нас,про,всего,них,какая,много,разве,сказала,три,эту,моя,впрочем,хорошо,свою,этой,перед,иногда,лучше,чуть,том,нельзя,такой,им,более,всегда,конечно,всю,между,меня,мне,мной,мною,ты,тебя,тебе,тобой,тобою,он,его,ему,им,него,нему,ним,она,ее,эи,ею,нее,нэи,нею,оно,его,ему,им,него,нему,ним,мы,нас,нам,нами,вы,вас,вам,вами,они,их,им,ими,них,ним,ними,себя,себе,собой,собою,этот,эта,это,эти,этого,эты,это,эти,этого,этой,этого,этих,этому,этой,этому,этим,этим,этой,этим,этою,этими,этом,этой,этом,этих,тот,та,то,те,того,ту,то,те,того,той,того,тех,тому,той,тому,тем,тем,той,тем,тою,теми,том,той,том,тех,весь,весь,вся,все,все,всего,всю,все,все,всего,всей,всего,всех,всему,всей,всему,всем,всем,всей,всем,всею,всеми,всем,всей,всем,всех,сам,сама,само,сами,самого,саму,само,самих,самого,самой,самого,самих,самому,самой,самому,самим,самим,самой,самим,самою,самими,самом,самой,самом,самих,быть,бы,буд,быв,есть,суть,име,дел,мог,мож,мочь,уме,хоч,хот,долж,можн,нужн,нельзя,Будут,люб,рич,ита,ладимир,наде,письм,нын,де,гост,пасиб,даст,ладимир,каков,арт,полн,Оссийск,каков,яв,видет,ладимир,будут,люб,рич,ос,арт,будут,люб,рич,будут,люб,рич,де,пасиб,ладимир,нын,нужд,рич,ита,яв,полн,лед,оссийск,будут,люб,рич,видет,Руз,важа,проч,будут,люб,рич,яв,едерац,будут,люб,рич,горазд,наде,одн,де,дин,ладимир,будут,люб,рич,ожн,рич,люб,будут,проч,знал,будут,рми,ладимирович,каков,видим,видим,ладимирович,рми,давн,рми,важа,леж,случ,знан,ладимирович,сег,рамк,рми,давн,ладимирович,рми,един,оп,случ,знан,ладимирович,видим,рми,дат,видим,знан,леж,ладимирович,ладимирович,видим,дат,заметн,ладимирович,онечн,сто,отмеч,личн,реч,основан,господин,парт,должн,видел,кром,ольк,рот,тат,кром,дава,вообщ,уш,трет,ком,реч,п,идт,ез,каза,пят,сет,ком,иха,ед,двойн,господ,сюд,мор,сет,руг,понятн,ред,видел,ита,ез,пят,шест,двойн,ред,ольк,сто,ед,отмеч,личн,реч,тип,онечн,основн,ведет,счет,месяц,собра,уш,трет,ез,каза,сет,тат,кром,дава,вообщ,сет,круг,п

Stop-stems:

будут,люб,рич,ита,ладимир,наде,письм,нын,де,гост,пасиб,даст,ладимир,каков,арт,полн,Оссийск,каков,яв,видет,ладимир,будут,люб,рич,ос,арт,будут,люб,рич,будут,люб,рич,де,пасиб,ладимир,нын,нужд,рич,ита,яв,полн,лед,оссийск,будут,люб,рич,видет,Руз,важа,проч,будут,люб,рич,яв,едерац,будут,люб,рич,горазд,наде,одн,де,дин,ладимир,будут,люб,рич,ожн,рич,люб,будут,проч,знал,будут,рми,ладимирович,каков,видим,видим,ладимирович,рми,давн,рми,важа,леж,случ,знан,ладимирович,сег,рамк,рми,давн,ладимирович,рми,един,оп,случ,знан,ладимирович,видим,рми,дат,видим,знан,леж,ладимирович,ладимирович,видим,дат,заметн,ладимирович,онечн,сто,отмеч,личн,реч,основан,господин,парт,должн,видел,кром,ольк,рот,тат,кром,дава,вообщ,уш,трет,ком,реч,п,идт,ез,каза,пят,сет,ком,иха,ед,двойн,господ,сюд,мор,сет,руг,понятн,ред,видел,ита,ез,пят,шест,двойн,ред,ольк,сто,ед,отмеч,личн,реч,тип,онечн,основн,ведет,счет,месяц,собра,уш,трет,ез,каза,сет,тат,кром,дава,вообщ,сет,круг,п,воениздат,ефимович,пыт,седьм,политолог,гот,миров,воз,едв,пыт,восем,явл

A small number of documents were removed because they contained no word counts. The Mutual Information Criterion was implemented separately in R.

After removing stems and creating the document-by-term matrix **X** (*d* x *t*), we begin by weighting the terms for maximum effect. We use two different criteria: the widely used Term Frequency by Inverse Document Frequency (*tf**idf*) and a Mutual Information Criterion. Both criteria are designed to weight words more highly when they are relatively rare in the corpus as a whole. *tf**idf* is a standard in the computational linguistics literature (Manning & Schütze 2003) and has been defended on theoretical grounds (Papineni 2001). The mutual information criterion is advocated by Hillard(2003) and includes a normalizing element. The formulas for each are as follows:

$$tf * idf = \mathbf{X}_{ij} * \log\frac{d}{d_{j>1}}$$

$$\mathrm{mutinf}_{ij} = \log\frac{p(w,t)}{p(w)p(t)} = \log\frac{p(w|t)}{p(w)}$$

where documents are indexed by **i** = 1...d Docs and terms are indexed by **j** = 1...t terms. After weighting the term counts we still have a term matrix that has dimensions (*d x t*). For dimensionality reduction we employ Latent Semantic Analysis (LSA). LSA was explored for document indexing as a way of reducing the feature space and linking words that coexist in the document as a whole (Deerwester et. al 1990), and has become a standard tool of topic models in the computational linguistics community (Manning & Schütze 2003, Manning et. al 2008, Tzoukerman et al. 2003).

LSA uses singular value decomposition (SVD) to map the weighted count matrix into a lower dimensional space by combining associated words into single features.[1] This feature reduction has the added benefit of addressing issues of synonymy. LSA allows two documents which may not share a word in common to be mapped onto a similar feature space because the words coexist in the corpus as a whole. For example, while one document may have the word "talk" and the other has the word "speak", LSA would most likely group them into the same feature space. We used two separate matrix reductions, one where we reduced to 100 features, and the other where we reduced to 500 features.

[1] This is similar to Principle Component Analysis (PCA) but does not require a singular matrix. Using SVD on the transpose of the weighted count matrix $\mathbf{X}^T$(*t x d*) yields matrices **U**(*t x m*), **D**(*m x m*) and $\mathbf{V}^T$(*m x d*), where **D** is a diagonal matrix whose singular values descend. By selecting the top *k* values, where $k < m$ and zeroing out the other values to create **D***, $\mathbf{UD^*V^T}$ becomes the least squares approximation of $\mathbf{X}^T$. With that intuition, we can see that $\mathbf{DV^T}$ is the matrix $\mathbf{X}^T$ projected into a lower dimensional space. By taking the transpose, we now have a matrix **X*** (*d x f)* where *f* is a number of features which is less than *t*. Similar feature reduction is completed in the Expressed Agenda Model using Latent Dirichlet Allocation, which is analogous to (although not the same as) probabilistic LSA. Notation conventions for the decomposition matrices follow the R notation for Singular Value Decomposition.

The described Latent Semantic Analysis was generated through use of the Singular Value Decomposition routine in the **`base`** R package. The _tf*idf_ matrix was reduced to the top k=100 values. The Mutual Information matrix was reduced to the top k=500 values. These values were chosen empirically by using cross-validation to check accuracy.

Table 1: Stems by Topic

	Russian Stems[2]	Translations	Category
Topic 1	Россия	Russia	*Interventionist*
	страст[ь]	passion	
	свободн[о]	free	
	сыгра[ть]	play	
	полиц[ия]	police (noun)	
	полицейск[ие]	police (adj.)	
	готов[ность]	readiness	
	этническ[ие]	ethnic	
	кпрф	CPRF (Communists)	
	межнациональн[ый]	interethnic	
Topic 2	войт[и]	enter	*Realpolitik*
	армен[ия]	Armenia	
	годовщин[а]	anniversary	
	военнослужа[щие]	military servicemen	
	войсков	forces	
	боеготов[ность]	military readiness	
	[за]служен[ный]	deserved	
	официальн[ый]	official	
	частн[ый]	private	
	возведен[ный]	elevated	
Topic 3	боеготов[ность]	military readiness	*Realpolitik*
	срочн[о]	immediately	
	войт[и]	enter	
	действительн[о]	really	
	управля[ть]	control	
	противовес	counterweight	
	задел	backup	
	сильн[ый]	strong	
	опережа[ть]	surpass	
	систематическ[и]	systematically	
Topic 4	военнослужа[щие]	military servicemen	*Realpolitik*
	воплот	bulwark	
	сильн[ый]	strong	
	находя[сь]	be situated	
	готов[ность]	readiness	
	безработиц[а]	unemployment	
	джордж	George	
	науч[ное]	scientific	
	Войсков	forces	
	москв[а]	Moscow	

C.3 UNSUPERVISED MODEL

The Expressed Agenda model was provided by Justin Grimmer. Upon publication of Grimmer (2009) the package will be released for R under the CRAN server. In order to determine the appropriate number of clusters we iterated over various values of k from 3 clusters to 15. We

[2] Prefixes and suffixes and translated words are illustrative and reflect one potential variation for “full” words expanded from the stems. In most cases, multiple variations are possible.

finally settled on k=6 at which point documents were labeled according to the top 20 words in the mutual information matrix. Table 1 provides the top 10 mutual information stems for the four labeled topics. The other two clusters were dropped as irrelevant (coded NA) and were removed from the model.

C.4 SUPERVISED MODELS

We employ four different machine learning approaches in our ensemble classifier, two of which are ensemble classifiers in their own right: K-Nearest Neighbor, Adaboost.M1 algorithm, Random Forest and Support Vector Machine. Each classifier is discussed in turn and Figure 1 shows the entire system.

K-Nearest Neighbor proceeds from an intuitive assumption that an unknown case can be classified according to its k nearest neighbors in the pa-rameter space (in this case using Minkowski distance). We selected k=5 and chose an un-weighted version of the model, where all five neighbors were treated equally (Hechenbiechler & Schliep 2004).

The Adaboost.M1 algorithm is a relatively recent discovery in machine learning, generating considerable interest following its introduction in the late 1990s (Polikar 2006). Adaboost.M1 uses several instances of a WeakLearner (in this implementation a classification tree) to generate hypotheses using data randomly drawn from the training distribution. The distribution is then iteratively updated to include instances misclassified by the first algorithm. This proceeds until a weighted majority vote occurs, which yields the final classification. Freund and Schapire (1997) provide a theoretical defense of the empirical success of the classifier, showing that as long as the error is less than .5 on each instance, the total error is bounded at the top and should decrease asymptotically with each iteration.

The Random Forest is similar to the Adaboost algorithm but uses a different approach. It is also an ensemble technique and uses classification trees as its sub-component. Rather than iteratively training near examples missed by the classifier previously, the trees are grown using bootstrapped versions of the data and by choosing k nodes for which to search for a split. This introduces random perturbations into the data which generate different results in each tree and prevents over-fitting, a common problem with decision trees (Breiman 1999, Breiman 2001a, Breiman 2001b).

Support Vector Machine (SVM) is easily the most popular machine learning algorithm in political science due to its easy implementation and broad utility compared to other techniques (Yang & Liu 1999). SVM fits a hyper-plane to the feature space, which separates two categories of points from each other and maximizes the marginal distance between the nearest points and the surface. A cost function determines the penalization of the soft-margins when inevitably all the points don't fall on one side of the plane. The logistical difficulty with SVM is that it is designed for dichotomous classification. We deal with this issue by using a common solution, running the SVM three times with each classification pitted against both others (NA versus ["Activist" + "Conservative"], "Activist" versus ["Conservative" + NA] etc.) and the classification with the greatest marginal distance from the hyper-plane is deemed the classifier's choice.

The first three algorithms are trained on the LSA-reduced tf*idf feature matrix (d x 100), and the SVM is trained on the LSA-reduced mutual information feature matrix (d x 500). The difference in feature sets and algorithms provides a diverse approach, beneficial given that no single classifier is superior for all classification problems.

The code for model estimation in each supervised learning technique is given below. Reference to the R package help file will illuminate the significance of chosen parameters and defaults:[3]

K-Nearest Neighbor

```
results <- kknn(TRAININGSET_VALUE ~., train = x, test = xtest, k = 5, kernel =
    "rectangular")
```

Adaboost.M1

```
results <- adaboost.M1(TRAININGSET_VALUE ~ ., data = x[-sample,], boos = TRUE, mfinal =
    100, coeflearn = ‘Breiman’, minsplit = 5, cp = 0.01, maxdepth = 18)
```

Support Vector Machine

```
results <- svmpath(x,NAvalue, epsilon=1e-6, lambda=0.1)
    predNA <- predict(results, xtest, lambda=1)
results <- svmpath(x,Avalue, epsilon=1e-6, lambda=0.1)
    predA <- predict(results, xtest, lambda=1)
results <- svmpath(x,Cvalue, epsilon=1e-6, lambda=0.1)
    predC <- predict(results, xtest, lambda=1)
```

Random Forest

```
results <- randomForest(x=x, y=y)
RFpredictions <- predict(results, data, type="response")
```

In order to effectively weight the predictions from the classifiers, we obtain accuracy measures by cross-validation for each algorithm. Using our set of 300 coded documents, we randomly sample 275, train the algorithm and then test on the out of sample 25 documents. We repeat this simulation 10,000 times for each algorithm to attain out-of-sample accuracy results.

We then weight the classifiers predictions according to the following formula where *p* is the number of correct out of sample predictions divided by the number of total out of sample predictions:

$$w_{class} = \log \frac{p}{1-p}$$

Given these weights, we simulate the training of each classifier on the same sample of 275, apply weights to the predictions and attain a final prediction and then verify against the out of sample 25. We repeat this process 10,000 times to produce the accuracy rates for the ensemble system as a whole. Accuracies for the individual classifiers and overall are given in the table below. The

[3] Note that the full R-Code and replication instructions are available in Appendix H. Code is provided here mainly as a reference for the critical specifications.

accuracy of the system is lower than any individual classifier, but not dramatically. Nonetheless, the ensemble approach guards against the idiosyncrasies of any given classifier.

Accuracy Rates

Classifier	Out-Of-Sample Error Rates
K-Nearest Neighbor	37.81%
Adaboost M.1	35.2%
Random Forest	34.36%
Support Vector Machine	48.08%
Ensemble	33.97%

Uncertainty

We use the estimated weights to incorporate the uncertainty of the classification procedure into our analysis, a form of uncertainty which is often discarded. We infer probabilities of categorization from the individual weights using the following formula:

$$p(Cat_i) = \frac{w_1 * Class1_i + w_2 * Class2_i + w_3 * Class3_i + w_4 * Class4_i}{w_1 + w_2 + w_3 + w_4}$$

where $ClassN_i$ is the dichotomous prediction of Classifier *N* for category *i*. We then redraw the data 10,000 times using a distribution defined by the probabilities for each class and observation. We then repeat the models using the new draws of the data. The mean of the coefficients is used as the coefficient of the estimate, and the standard deviation of the coefficients is used as the standard error, which incorporates the uncertainty in the classification process.

The probabilities of accuracy are measured as the probability of a correct classification on 10,000 cross-validation runs training on a random selection of 275 documents and holding back 25 documents. Probabilities are based only on this collection of 25 out-of-sample classification rates.

Estimates on replication will be imprecise due to the use of simulation in the text. However due to the large number of cross-validation runs, the estimates should converge. The process is extremely computationally intensive and over 10,000 runs the code should be expected to take several hours to run even on a newer machine.

APPENDIX D: UNCERTAINTY

Whether in automated or manual document analysis, error rates or inter-coder reliability rates are rarely taken into account at the level of estimation. Hopkins and King (2008) present a supervised document classification model called ReadMe that helps to ameliorate this deficiency by removing the error prone step of individual document classification. The algorithm, in contrast to supervised methods in the computer science literature, computes the distribution of document classifications rather than to classify each individual document. Unfortunately, we need individual level data, both for the analysis of our first three hypotheses and because several of the months in the time series have too few documents for an adequate distribution.

We can understand this procedure in two ways which are theoretically distinct but in practice the same. In the first understanding, the entire DV can be treated as missing, with each sample being a multiple imputation estimate (King et al. 2000). The second understanding is that we are performing a weighting of observations commensurate with our understood uncertainty. In either case, the procedure is the same. The one important difference is that under the construct of multiple imputation, we can include all of our meta-data into the classification process (the author, affiliation, date etc.); to fail to include any data that will eventually be used in the analysis from the imputation process can result in bias. We tested the inclusion of the meta-data and there was no difference in the analyses.

REFERENCES

Alfaro, E., Gamez, M. and Garcia, N. 2007: “Multiclass corporate failure prediction by Adaboost.M1.” *International Advances in Economic Research*, Vol 13, 3, pp. 301–312.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research*. 3: 993-1022.

Blei, David and John Lafferty. 2006. “Dynamic Topic Models.” *Proceedings of the 23rd International Conference on Machine Learning* 23.

Breiman, Leo. 2001a, “Random Forests.” *Machine Learning*. 45(1), 5-32.

Breiman, Leo. 2001b. “Statistical Modeling: The Two Cultures.” *Statistical Science*. 16:3, 199-231.

Breiman, Leo. 2002, “Manual On Setting Up, Using, And Understanding Random Forests V3.1”, Research Note.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman. 1990. “Indexing by latent semantic analysis.” *Journal of the American Society for Information Science*, 41(6), 391-407.

Embretson, Susan E. and Steven P. Reise. 2000. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Freund, Y. and Schapire, R.E. 1996: “Experiments with a New Boosting Algorithm”. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148–156, Morgan Kaufmann.

Fox, John. 2008. car: Companion to Applied Regression. R package version1.2-9.

Freund, Y. and Schapire, R.E. 1997. “Decision-theoretic generalization of on-line learning and an application to boosting” *Journal of Computer and System Sciences*, vol. 55 no. 1, pp. 119-139.

Gelman, Andrew, Gary King and Chuanhai Liu. 1999. “Not Asked and Not Answered: Multiple Imputation for Multiple Surveys.” *Journal of the American Statistical Association*, Vol. 93, No. 443: Pp. 846-857.

Godbole, Namrata, Manjunath Srinivasaiah and Steven Skiena. 2007. “Large-Scale Sentiment Analysis for News and Blogs.” Proceedings of International Conference on Weblogs and Social Media.

Grimmer, Justin. 2008. “A Bayesian Hierarchical Topic Model for Political Texts: Supplemental Appendix.” Working Paper.

Grimmer, Justin. 2009. “A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases.” Working Paper.

Hastie, Trevor, Saharon Rosset, Robert Tibshirani and Ji Zhu. 2004. “The Entire Regularization Path for the Support Vector Machine.” *Journal of Machine Learning Research*. Vol 5, 1391-1415.

Trevor Hastie 2006. svmpath: svmpath: the SVM Path algorithm. R package version 0.92. http://www-stat.stanford.edu/~hastie/Papers/svmpath.pdf

Hillard, Dustin. 2003. “Topic Classification for Conversational Speech using Support Vector Machine and Latent Semantic Analysis.” Working Paper.

Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2007. “An Active Learning Framework for Classifying Political Text.” Midwest Political Science Association 65th Annual National Conference.

Hillard, Dustin, Stephen Purpura and John Wilkerson. 2008. “Computer Assisted Topic Classification for Mixed Methods Social Science Research.” *Journal of Information Technology and Policy*. 4(4).

Hopkins, Daniel and Gary King. 2008. “A Method of Automated Nonparametric Content Analysis for Social Science.” Working Paper. Previously presented at Midwest Political Science Association and the Society for Political Methodology.

Imai, Kosuke, Gary King, and Oliva Lau. 2007. ”negbin: Negative Binomial Regression for Event Count Dependent Variables” in Kosuke Imai, Gary King, and Olivia Lau, ”Zelig: Everyone’s Statistical Software,”http://gking.harvard.edu/zelig

Imai,Kosuke, Gary King, and Olivia Lau. 2007. “Zelig: Everyone’s Statistical Software,” http://GKing.harvard.edu/zelig.

Imai, Kosuke, Gary King, and Olivia Lau. 2008. “Toward A Common Framework for Statistical Analysis and Development.” *Journal of Computational and Graphical Statistics*, Vol. 17, No. 4 (December), pp. 892-913.

Joachims, T. 1998. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning (ECML).

Liaw, A. and M. Wiener. 2002. Classification and Regression by randomForest. R News 2(3), 18--22.

Manning, Christopher D. and Hinrich Schütze. 2003. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Manning, Chrisopher D., Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Monroe, Burt, Michael Colaresi and Kevin Quinn. 2009. “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict.” Forthcoming in *Political Analysis*.

Papineni, Kishore. 2001. “Why inverse document frequency?” NAACL Proceedings, 25-32.

Polikar, Robi. 2006. “Ensemble Based Systems in Decision Making.” *IEEE Circuits and Systems Magazine*.21-45.

Porter, M. F. 1980. “An algorithm for suffix stripping.” Program 14(3):130–137.

Purpura, Stephen and Dustin Hillard. 2006. “Automated Classification of Congressional Legislation.” The 7th Annual International Conference on Digital Government Research. May 21-24, San Diego, CA.

Purpora, Stephen, Dustin Hillard and Philip Howard. 2006a. “A Comparative Study of Human Coding and Context Analysis against Support Vector Machines (SVM) to Differentiate Campaign Emails by Party and Issues.” Working Paper.

Quinn, Kevin, Burt L. Monroe, Michael Colaresi, Michael Crespin, and Dragomir Radev. 2006. “How To Analyze Political Attention With Minimal Assumptions And Costs.” Society for Political Methodology, Davis, CA.

Schliep, Klaus and Klaus Hechenbichler (2008). kknn: Weighted k-Nearest Neighbors. R package version 1.0-6.

Schrodt, Philip, Palmer, Glenn and Hatipoglu, Mehmet. 2008. “Automated Detection of Militarized Interstate Disputes Using Document Classification Algorithms” Paper presented at the annual meeting of the APSA 2008 Annual Meeting, Hynes Convention Center, Boston, Massachusetts, Aug 28, 2008 Online <WEBMAIL/PDF>. 2009-02-16 <http://www.allacademic.com/meta/p278383_index.html>

Shulman, Stuart. 2008. “Editor’s Introduction.” *Journal of Information Technology & Politics*. 5(4), 353-354.

Snowball Stemmer. Software. http://snowball.tartarus.org/

Tzoukerman, Evelyne, Judith Klavans and Tomek Strzalkowski. 2003. “Information Retrieval.” Oxford Handbook of Computational Linguistics. Ed. Ruslan Mitkov.

Vapnic, V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Venables, W. N. and Ripley, B. D. 2002. *Modern Applied Statistics with S*. Fourth Edition. Springer, New York.