Appendix A
REUTERS-21578

A.1 Introduction

This chapter is a condensed version of the information available about the Reuters-21578 collection.

Reuters-21578 text categorization test collection is a resource for research in information retrieval, machine learning, and other corpus-based research.

The copyright for the text of newswire articles and Reuters annotations in the Reuters-21578 collection resides with Reuters Ltd. and Carnegie Group, Inc. that have agreed to allow the free distribution of this data for research purposes only.

The Reuters-21578, Distribution 1.0 test collection is available from http://www.daviddlewis.com/resources/testcollections/reuters21578

A.2 History

The documents in the Reuters-21578 collection appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd. (Sam Dobbins, Mike Topliss, Steve Weinstein) and Carnegie Group, Inc. (Peggy Andersen, Monica Cellio, Phil Hayes, Laura Knecht, Irene Nirenburg) in 1987.

In 1990, the documents were made available by Reuters and CGI for research purposes to the Information Retrieval Laboratory (W. Bruce Croft, Director) of the Computer and Information Science Department at the University of Massachusetts at Amherst. Formatting of the documents and production of associated data files was done in 1990 by David D. Lewis and Stephen Harding at the Information Retrieval Laboratory.

Further formatting and data file production was done in 1991 and 1992 by David D. Lewis and Peter Shoemaker at the Center for Information and Language Studies, University of Chicago. This version of the data was made available for anonymous FTP as "Reuters-22173, Distribution 1.0" in January 1993. From 1993 through 1996, Distribution 1.0 was hosted at a succession of FTP sites maintained by the

Center for Intelligent Information Retrieval (W. Bruce Croft, Director) of the Computer Science Department at the University of Massachusetts at Amherst.

At the ACM SIGIR '96 conference in August, 1996, a group of text categorization researchers discussed how published results on Reuters-22173 could be made more comparable across studies. It was decided that a new version of collection should be produced with less ambiguous formatting, and including documentation carefully spelling out standard methods of using the collection. The opportunity would also be used to correct a variety of typographical and other errors in the categorization and formatting of the collection.

Steve Finch and David D. Lewis did this cleanup of the collection September through November of 1996, relying heavily on Finch's SGML-tagged version of the collection from an earlier study. One result of the re-examination of the collection was the removal of 595 documents which were exact duplicates (based on identity of timestamps down to the second) of other documents in the collection. The new collection therefore has only 21,578 documents, and thus is called the Reuters-21578 collection.

A.3 Formatting

The Reuters-21578 collection is distributed in 22 files. Each of the first 21 files (`reut2-000.sgm` through `reut2-020.sgm`) contain 1000 documents, while the last (`reut2-021.sgm`) contains 578 documents.

The files are in SGML format. Rather than going into the details of the SGML language, it is described how the SGML tags are used to divide each file, and each document, into sections.

Each of the 22 files begins with a document type declaration line:
`<!DOCTYPE lewis SYSTEM "lewis.DTD">`

A.4 The REUTERS Tag

Each article starts with an "open tag" of the form:
`<REUTERS TOPICS=?? LEWISSPLIT=?? CGISPLIT=?? OLDID=?? NEWID=??>`
where the `??` are filled in an appropriate fashion. Each article ends with a "close tag" of the form:`</REUTERS>`

In all cases the `<REUTERS>` and `</REUTERS>` tags are the only items on their line.

Each `REUTERS` tag contains explicit specifications of the values of five attributes, `TOPICS, LEWISSPLIT, CGISPLIT, OLDID,` and `NEWID`. These attributes are meant to identify documents and groups of documents, and have the following meanings:

1. TOPICS: The possible values are YES, NO and BYPASS:
   a. YES: indicates that in the original data there was at least one entry in the TOPICS fields;
   b. NO: indicates that in the original data the story had no entries in the TOPICS field;
   c. BYPASS: indicates that in the original data the story was marked with the string BYPASS (or a typographical variant on that string).

   This poorly-named attribute unfortunately is the subject of much confusion. It is meant to indicate whether or not the document had TOPICS categories in the raw Reuters-22173 dataset. The sole use of this attribute is to defining training set splits similar to those used in previous research. (See the section on training set splits.) The TOPICS attribute does not indicate anything about whether or not the Reuters-21578 document has any TOPICS categories. That can be determined by actually looking at the TOPICS field. A story with TOPICS="YES" can have no TOPICS categories, and a story with TOPICS="NO" can have TOPICS categories.

   A reasonable (though not certain) assumption is that for all TOPICS="YES" stories the indexer at least thought about whether the story belonged to a valid TOPICS category. Thus, the TOPICS="YES" stories with no topics can reasonably be considered negative examples for all 135 valid TOPICS categories.

   TOPICS="NO" stories are more problematic in their interpretation. Some of them presumedly result because the indexer made an explicit decision that they did not belong to any of the 135 valid TOPICS categories. However, there are many cases where it is clear that a story should belong to one or more TOPICS categories, but for some reason the category was not assigned. There appear to be certain time intervals where large numbers of such stories are concentrated, suggesting that some parts of the data set were simply not indexed, or not indexed for some categories or category sets. Also, in a few cases, the indexer clearly meant to assign TOPICS categories, but put them in the wrong field. These cases have been corrected in the Reuters-21578 data, yielding stories that have TOPICS categories, but where TOPICS="NO", because the the category was not assigned in the raw version of the data.

   "BYPASS" stories clearly were not indexed, and so are useful only for general distributional information on the language used in the documents.
2. LEWISSPLIT: The possible values are TRAINING, TEST, and NOT-USED. TRAINING indicates it was used in the training set in the experiments reported in [66, 67, 68, 71]. TEST indicates it was used in the test set for those experiments, and NOT-USED means it was not used in those experiments.
3. CGISPLIT: The possible values are TRAINING-SET and PUBLISHED-TESTSET indicating whether the document was in the training set or the test set for the experiments reported in [42, 43].
4. OLDID: The identification number (ID) the story had in the Reuters-22173 collection.

5. NEWID: The identification number (ID) the story has in the Reuters-21578, Distribution 1.0 collection. These IDs are assigned to the stories in chronological order.

In addition, some REUTERS tags have a sixth attribute, CSECS, which can be ignored.

The use of these attributes is critical to allowing comparability between different studies with the collection.

A.5 Document-Internal Tags

Just as the <REUTERS> and </REUTERS> tags serve to delimit documents within a file, other tags are used to delimit elements within a document. These are discussed in the order in which they typically appear, though the exact order should not be relied upon in processing. In some cases, additional tags occur within an element delimited by these top level document-internal tags. These are discussed in this section as well.

It is specified below whether each open/close tag pair is used exactly once (ONCE) per a story, or a variable (VARIABLE) number of times (possibly zero). In many cases the start tag of a pair appears only at the beginning of a line, with the corresponding end tag always appearing at the end of the same line. When this is the case, it is indicated it with the notation "SAMELINE" below, as an aid to those processing the files without SGML tools.

1. <DATE>, </DATE> [ONCE, SAMELINE]: Encloses the date and time of the document, possibly followed by some non-date noise material.
2. <MKNOTE>, </MKNOTE> [VARIABLE]: Notes on certain hand corrections that were done to the original Reuters corpus by Steve Finch.
3. <TOPICS>, </TOPICS> [ONCE, SAMELINE]: Encloses the list of TOPICS categories, if any, for the document. If TOPICS categories are present, each will be delimited by the tags <D> and </D>.
4. <PLACES>, </PLACES> [ONCE, SAMELINE]: Same as <TOPICS> but for PLACES categories.
5. <PEOPLE>, </PEOPLE> [ONCE, SAMELINE]: Same as <TOPICS> but for PEOPLE categories.
6. <ORGS>, </ORGS> [ONCE, SAMELINE]: Same as <TOPICS> but for ORGS categories.
7. <EXCHANGES>, </EXCHANGES> [ONCE, SAMELINE]: Same as <TOPICS> but for EXCHANGES categories.
8. <COMPANIES>, </COMPANIES> [ONCE, SAMELINE]: These tags always appear adjacent to each other, since there are no COMPANIES categories assigned in the collection.
9. <UNKNOWN>, </UNKNOWN> [VARIABLE]: These tags bracket control characters and other noisy and/or somewhat mysterious material in the Reuters stories.

10. <TEXT>, </TEXT> [ONCE]: There was an attempt to delimit all the textual material of each story between a pair of these tags. Some control characters and other "junk" material may also be included. The whitespace structure of the text has been preserved. The <TEXT> tag has the following attributes:

    a. TYPE: This has one of three values: NORM, BRIEF, and UNPROC. NORM is the default value and indicates that the text of the story had a normal structure. In this case the TEXT tag appears simply as <TEXT>. The tag appears as <TEXT TYPE="BRIEF"> when the story is a short one or two line note. The tags appears as <TEXT TYPE="UNPROC"> when the format of the story is unusual in some fashion that limited our ability to further structure it.

       The following tags optionally delimit elements inside the TEXT element. Not all stories will have these tags:

    b. <AUTHOR>, </AUTHOR>: Author of the story.
    c. <DATELINE>, </DATELINE>: Location the story originated from, and day of the year.
    d. <TITLE>, </TITLE>: Title of the story. An attempt to capture the text of stories with TYPE="BRIEF" within a <TITLE> element.
    e. <BODY>, </BODY>: The main text of the story.

A.6 Categories

A test collection for text categorization contains, at minimum, a set of texts and, for each text, a specification of what categories that text belongs to. For the Reuters-21578 collection the documents are Reuters newswire stories, and the categories are five different sets of content related categories, as shown in table A.1. For each document, a human indexer decided which categories from which sets that document belonged to.

The TOPICS categories are economic subject categories. Examples include "coconut", "gold", "inventories", and "money-supply". This set of categories is the one that has been used in almost all previous research with the Reuters data.

The EXCHANGES, ORGS, PEOPLE, and PLACES categories correspond to named entities of the specified type. Examples include "nasdaq" (EXCHANGES),

Table A.1 Category types in Reuters-21578.

Type	Number	+ 1 occurrence	+20 occurrences
EXCHANGES	39	32	7
ORGS	56	32	9
PEOPLE	267	114	15
PLACES	175	147	60
TOPICS	135	120	57

”gatt” (ORGS), ”perez-de-cuellar” (PEOPLE), and ”australia” (PLACES). Typically a document assigned to a category from one of these sets explicitly includes some form of the category name in the document’s text. (Something which is usually not true for TOPICS categories.)

Reuters-21578, Distribution 1.0 includes five files (all-exchanges-strings.lc.txt, all-orgs-strings.lc.txt, all-people-strings.lc.txt, all-places-strings.lc.txt, and all-topics-strings.lc.txt) which list the names of all legal categories in each set. A sixth file, cat-descriptions_120396.txt gives some additional information on the category sets.

Note that a sixth category field, COMPANIES, was present in the original Reuters materials distributed by Carnegie Group, but no company information was actually included in these fields. In the Reuters-21578 collection this field is always empty.

In the table above it can be seen how many categories appear in at least 1 of the 21,578 documents in the collection, and how many appear at least 20 of the documents. Many categories appear in no documents, but researchers are encouraged to include these categories when evaluating the effectiveness of their categorization system.

Additional details of the documents, categories, and corpus preparation process appear in [67], and at greater length in [66].

A.7 Using Reuters-21578 for Text Categorization Research

In testing a method for text categorization it is important that knowledge of the nature of the test data not unduly influence the development of the system, or the performance obtained will be unrealistically high. One way of dealing with this is to divide a set of data into two subsets: a training set and a test set. An experimenter then develops a categorization system by automated training on the training set only, and/or by human knowledge engineering based on examination of the training set only. The categorization system is then tested on the previously unexamined test set.

Effectiveness results can only be compared between studies that the same training and test set (or that use cross-validation procedures). One problem with the Reuters-22173 collection was that the ambiguity of formatting and annotation led different researchers to use different training/test divisions. This was particularly problematic when researchers attempted to remove documents that ”had no TOPICS”, as there were several definitions of what this meant.

To eliminate these ambiguities from the Reuters-21578 collection, it is defined exactly which articles are in each of the recommended training sets and test sets by specifying the values those articles will have on the TOPICS, LEWISSPLIT, and CGISPLIT attributes of the REUTERS tags. It is strongly encouraged that all studies on Reuters-21578 use one of the following training test divisions (or use multiple random splits, e.g. cross-validation).

A.7.1 The Modified Lewis ("ModLewis") Split

This split replaces the 14704/6746 split (723 unused) of the Reuters-22173 collection, which was used in [66, 67, 68, 71].

Table A.2 presents the splitting procedure. If `LEWISSPLIT` is `NOT-USED` or `TOPICS` is `BYPASS`, a document is not used.

The duplicate documents removed in forming Reuters-21578 are of course not present. The documents with TOPICS="BYPASS" are not used, since subsequent analysis strongly indicates that they were not categorized by the indexers. The 1,765 unused documents should not be tested on and should not be used for supervised learning. However, they may useful as additional information on the statistical distribution of words, phrases, and other features that might used to predict categories.

This split assigns documents from April 7, 1987 and before to the training set, and documents from April 8, 1987 and after to the test set. Table A.3 presents the number of documents used on ModLewis split.

Given the many changes in going from Reuters-22173 to Reuters-21578, including correction of many typographical errors in category labels, results on the ModLewis split cannot be compared with any published results on the Reuters-22173 collection.

Table A.2 ModLewis split.

Set	LEWISSPLIT	TOPICS
Train	TRAIN	YES ou NO
Test	TEST	Yes ou NO
Not Used	NOT-USED	-
Not Used	-	BYPASS

Table A.3 Used documents on ModLewis split.

Set	Documents
Train	13625
Test	6188
Nou used	1765

A.8 The Modified Apte ("ModApte") Split

This replaces the 10645/3672 split (7,856 not used) of the Reuters-22173 collection. It is an approximation to the training and test splits used in [6] and [7].

Table A.4 presents the split. As with the ModLewis, those documents removed in forming Reuters-21578 are not present, and BYPASS documents are not used.

The intent in [6] and [7] to use the Lewis split, but restrict it to documents with at least one `TOPICS` categories. However, but it was not clear exactly what

Table A.4 ModApte split.

Set	LEWISSPLIT	TOPICS
Train	TRAIN	YES
Test	TEST	YES
Not Used	NOT-USED	YES
Not Used	-	NO
Not Used	-	BYPASS

Apte, et al meant by having at least one TOPICS category (e.g. how was "bypass" treated, whether this was before or after any fixing of typographical errors, etc.). This interpretation is encoded in the TOPICS attribute.

As discussed above, some TOPICS="YES" stories have no TOPICS categories, and a few TOPICS="NO" stories have TOPICS categories. These facts are irrelevant to the definition of the split.

If you are using a learning algorithm that requires each training document to have at least TOPICS category, you can screen out the training documents with no TOPICS categories.

Please do NOT screen out any of the 3,299 documents - that will make your results incomparable with other studies.

As with ModLewis, it may be desirable to use the 8,676 unused documents for gathering statistical information about feature distribution.

As with ModLewis, this split assigns documents from April 7, 1987 and before to the training set, and documents from April 8, 1987 and after to the test set. The difference is that only documents with at least one TOPICS category are used. The rationale for this restriction is that while some documents lack TOPICS categories because no TOPICS apply (i.e. the document is a true negative example for all TOPICS categories), it appears that others simply were never assigned TOPICS categories by the indexers. (Unfortunately, the amount of time that has passed since the collection was created has made it difficult to establish exactly what went on during the indexing.)

Table A.5 presents the number of documents used on ModApte split.

Given the many changes in going from Reuters-22173 to Reuters-21578, including correction of many typographical errors in category labels, results on the ModApte split cannot be compared with any published results on the Reuters-22173 collection.

Table A.5 Used documents on ModApte split.

Set	Documents
Train	9603
Test	3299
Not used	8676

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="5552" NEWID="9">
<DATE>26-FEB-1987 15:17:11.20</DATE>
<TOPICS><D>earn</D></TOPICS>
<PLACES><D>usa</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;F
&#22;&#22;&#1;f0762&#31;reute
r f BC-CHAMPION-PRODUCTS-&lt;CH   02-26 0067</UNKNOWN>
<TEXT>&#2;
<TITLE>CHAMPION PRODUCTS &lt;CH> APPROVES STOCK SPLIT</TITLE>
<DATELINE>    ROCHESTER, N.Y., Feb 26 - </DATELINE><BODY>Champion
Products Inc said its
board of directors approved a two-for-one stock split of its
common shares for shareholders of record as of April 1, 1987.
    The company also said its board voted to recommend to
shareholders at the annual meeting April 23 an increase in the
authorized capital stock from five mln to 25 mln shares.
 Reuter
&#3;</BODY></TEXT>
</REUTERS>
```

Fig. A.1 Example of a Reuters-21578 document.

A.9 Stopwords

a; about; above; across; after; afterwards; again; against; all; almost; alone; along; already; also; although; always; am; among; amongst; amoungst; amount; an; and; another; any; anyhow; anyone; anything; anyway; anywhere; are; around; as; at; back; be; became; because; become; becomes; becoming; been; before; beforehand; behind; being; below; beside; besides; between; beyond; bill; both; bottom; but; by; call; can; cannot; cant; co; computer; con; could; couldnt; cry; de; describe; detail; do; done; down; due; during; each; eg; eight; either; eleven; else; elsewhere; empty; enough; etc; even; ever; every; everyone; everything; everywhere; except; few; fifteen; fify; fill; find; fire; first; five; for; former; formerly; forty; found; four; from; front; full; further; get; give; go; had; has; hasnt; have; he; hence; her; here; hereafter; hereby; herein; hereupon; hers; herself; him; himself; his; how; however; hundred; i; ie; if; in; inc; indeed; interest; into; is; it; its; itself; keep; last; latter; latterly; least; less; ltd; made; many; may; me; meanwhile; might; mill; mine; more; moreover; most; mostly; move; much; must; my; myself; name; namely; neither; never; nevertheless; next; nine; no; nobody; none; noone; nor; not; nothing; now; nowhere; of; off; often; on; once; one; only; onto; or; other; others; otherwise; our; ours; ourselves; out; over; own; part; per; perhaps; please; put; rather; re; same; see; seem; seemed; seeming; seems; serious; several; she; should; show; side; since; sincere; six; sixty; so; some; somehow; someone; something; sometime; sometimes; somewhere; still; such; system; take; ten; than; that; the; their; them; themselves;

then; thence; there; thereafter; thereby; therefore; therein; thereupon; these; they; thick; thin; third; this; those; though; three; through; throughout; thru; thus; to; together; too; top; toward; towards; twelve; twenty; two; un; under; until; up; upon; us; very; via; was; we; well; were; what; whatever; when; whence; whenever; where; whereafter; whereas; whereby; wherein; whereupon; wherever; whether; which; while; whither; who; whoever; whole; whom; whose; why; will; with; within; without; would; yet; you; your; yours; yourself; yourselves;

Appendix B
RCV1 - Reuters Corpus Volume I

B.1 Introduction

This chapter is a rather condensed version of the information available about the Reuters Corpus Volume I (RCV1) collection. RCV1 is an archive of over 800,000 manually categorized newswire stories using three category sets, that was recently made available by Reuters Ltd. for research purposes. Use of this data for research on text categorization requires a detailed understanding of the real world constraints under which the data was produced. RCV1 as distributed is simply a collection of newswire stories, not a test collection. It includes known errors in category assignment, provides lists of category descriptions that are not consistent with the categories assigned to articles, and lacks essential documentation on the intended semantics of category assignment [107]. Nevertheless it constitutes a valuable research tool for text, namely for categorization.

Reuters Ltd. has gone through significant restructuring since RCV1 was produced, and information that in a research setting would have been retained was therefore not recorded. In particular, no formal specification remains of the coding practices at the time the RCV1 data was produced. Fortunately, several researchers [107, 73] have examined the available information and, by combining related documentation and interviews with Reuters personnel, have largely reconstructed those aspects of coding relevant to text categorization research.

B.2 The Documents

Reuters Ltd. is the largest international text and television news agency. Its editorial division produces some 11,000 stories a day in 23 languages [73]. Stories are both distributed in real time and made available via online databases and other archival products.

The RCV1 dataset was created from one of those online databases. It consists of the English language stories produced by Reuters journalists between August 20, 1996, and August 19, 1997. A researcher can obtain the data on two CD-ROMs,

formatted in XML, by submiting a request to the National Institute of Science and Technology[1]. Figure B.1 shows an example story with some simplification of the markup for brevity, taken from [73].

RCV1 contains 35 times as many documents (806,791) as the popular Reuters-21578 collection (see Appendix A), making it one of the largest available text categorization test collection. Moreover, RCV1 is also more organized than previous collections. Each document is in a separate file and has a unique ID, ranging from 2286 to 810597 with some gaps. The ID order does not correspond to chronological order of the stories, but they have time stamps that give only the day, not the time, since the stories were taken from an archival database, not from the original stream sent out over the newswire.

Both text and metadata are formatted with XML, simplifying their use. As RCV1 was produced from an archival database, it has fewer alerts, corrections to previous stories, and other peculiarities.

RCV1 contains all or almost all stories of a particular type from an interval of one year. For temporal studies, this is a major advantage over Reuters-21578, which had uneven coverage of a fraction of a year.

The corpus has a limited number of duplicates, foreign language documents, and other similar issues, which can be problematic depending on the application at hand, but are comparable to to levels seen in operational settings.

B.3 The Categories

To aid retrieval from database products category codes from three sets (Topics, Industries, and Regions) were assigned to stories.

B.3.1 Topic Codes

In the experiments carried out in this thesis categories were taken from the topic codes, which were assigned to capture the major subjects of a story. They were organized in four hierarchical groups: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT(Markets).

There are 103 categories actually assigned to the data. In the second column of Table B.1 (Documents) the total number of positive documents for the ten most populated classes of RCV1 is given. For training, RCV1 defines 22,370 documents that should be used. For testing we selected the first 50,000 documents not used for training. Table B.1 also presents the effective number of positive documents used for training and testing.

One can see that this code set shows particular perspective on a data set. The RCV1 articles span a broad range of content, but the code set only emphasizes distinctions relevant to Reuters customers. For instance,there are three different Topic codes for corporate ownership changes, but all of science and technology is a single category (GSCI) [72].

[1] http://trec.nist.gov/data/reuters/reuters.html

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="2330" id="root" date="1996-08-20" xml:lang="en">
<title>USA: Tylan stock jumps; weighs sale of company.</title>
<headline>Tylan stock jumps; weighs sale of company.</headline>
<dateline>SAN DIEGO</dateline>
<text>
<p>The stock of Tylan General Inc. jumped Tuesday after the maker of
process-management equipment said it is exploring the sale of the
company and added that it has already received some inquiries from
potential buyers.</p>
<p>Tylan was up $2.50 to $12.75 in early trading on the Nasdaq market.</p>
<p>The company said it has set up a committee of directors to oversee
the sale and that Goldman, Sachs &amp;amp; Co. has been retained as its
financial adviser.</p>
</text>
<copyright>(c) Reuters Limited 1996</copyright>
<metadata>
<codes class="bip:countries:1.0">
<code code="USA"> </code>
</codes>
<codes class="bip:industries:1.0">
<code code="I34420"> </code>
</codes>
<codes class="bip:topics:1.0">
<code code="C15"> </code>
<code code="C152"> </code>
<code code="C18"> </code>
<code code="C181"> </code>
<code code="CCAT"> </code>
</codes>
<dc element="dc.publisher" value="Reuters Holdings Plc"/>
<dc element="dc.date.published" value="1996-08-20"/>
<dc element="dc.source" value="Reuters"/>
<dc element="dc.creator.location" value="SAN DIEGO"/>
<dc element="dc.creator.location.country.name" value="USA"/>
<dc element="dc.source" value="Reuters"/>
</metadata>
</newsitem>
```

Fig. B.1 Example of a RCV1 document.

Table B.1 Positive documents for RCV1 categories.

Category	Documents	Training	Testing
CCAT	381327	10416	23077
GCAT	239267	2050	5180
MCAT	204820	5154	11110
C15	151785	3122	7454
ECAT	119920	3162	7539
M14	85440	1799	4887
C151	81890	515	698
C152	73092	1088	435
GPOL	56878	1627	3756
M13	53634	1095	2613

B.3.1.1 Industry Codes

Industry codes were assigned based on types of businesses discussed in the story. They were grouped in 10 subhierarchies,such as I2(METALSANDMINERALS) and I5(CONSTRUCTION). The Industry codes make up the largest of the three code sets, supporting many fine distinctions.

B.3.1.2 Region Codes

Region codes included both geographic locations and economic/political groupings. No hierarchical taxonomy was defined.

B.3.2 Coding Policy

Explicit policies on code assignment presumedly increase consistency and usefulness of coding, though coming up with precise policies is difficult. Reuters guidance for coding included two broad policies, among others. In [73], these policies are described as:

1. Minimum Code Policy: Each story was required to have at least one Topic code and one Region code;
2. Hierarchy Policy: Coding was to assign the most specific appropriate codes from the Topic and Industry sets, as well as (usually automatically) all ancestors of those codes. In contrast to some coding systems, there was no limit on the number of codes with the same parent that could be applied.

B.4 Stopwords

For comparison purposes, the same set of stopwords used for Reuters-21578 (see Appendix A) was filtered out from RCV1.

References

1. Aas, K., Eikvil, L.: Text categorisation: A survey. Technical report, Norwegian Computing Center (1999)
2. Abdalhaq, B., Cortés, A., Margalef, T., Bianchini, G., Luque, E.: Between classical and ideal: Enhancing wildland fire prediction using cluster computing. Cluster Computing 9(3), 329–343 (2006)
3. Abernethy, J., Bach, F., Evgeniou, T., Vert, J.: A new approach to collaborative filtering: Operator estimation with spectral regularization. Journal of Machine Learning Research 10, 803–826 (2009)
4. Al-Ani, A., Deriche, M.: Feature selection using a mutual information based measure. In: International Conference on Pattern Recognition - ICPR 2002, vol. 4, pp. 82–85 (2002)
5. Androutsopoulos, I., Koutsias, J., Chandrinos, K., Spyropoulos, C.: An experimental comparison of naïve bayesian and keyword-based anti-spam filtering with personal e-mail messages. In: International Conference on Research and Development in InformationRetrieval - ACM SIGIR 2000, pp. 160–167 (2000)
6. Apté, C., Damerau, F., Weiss, S.: Automated learning of decision rules for text categorization. ACM Transactions for Information Systems 12, 233–251 (1994)
7. Apté, C., Damerau, F., Weiss, S.: Toward language independent automated learning of text categorization models. In: International Conference on Research and Development in InformationRetrieval- ACM SIGIR 1994, pp. 23–30 (1994)
8. Armano, G., Manconi, A., Vargiu, E.: A multiagent system for retrieving bioinformatics publications from web sources. IEEE Transactions on NanoBioscience 6(2), 104–109 (2007)
9. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, New York (1999)
10. Baker, L., McCallum, A.: Distributional clustering of words for text classification. In: International Conference on Research and Development in Information Retrieval - ACM SIGIR 1998, pp. 96–103 (1998)
11. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning 36(1–2), 105–139 (1999)
12. Benali, F., Ubeda, S., Legrand, V.: Collaborative approach to automatic classification of heterogeneous information security. In: Second International Conference on Emerging Security Information, Systems and Technologies, pp. 294–299 (2008)
13. Bishop, C.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)

14. Bishop, C., Tipping, M.: Bayesian Regression and Classification. In: Suykens, j., Horvath, G., Basu, S., Micchelli, C., Vandewalle, J. (eds.) Advances in Learning Theory: Methods, Models and Applications. NATO Science Series III: Computer and Systems Sciences, vol. 190, pp. 267–285. IOS Press, Amsterdam (2003)
15. Błażewicz, J., Ecker, K., Pesch, E., Schmidt, G., Weglarz, J.: Scheduling Computer and Manufacturing Processes. Springer, Heidelberg (2001)
16. Boser, B., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: ACM Annual Workshop on Computational Learning Theory - COLT 1992, pp. 144–152 (1992)
17. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
18. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Computer Networks and ISDN Systems, pp. 107–117 (1998)
19. Chin, O., Kulathuramaiyer, N., Yeo, A.: Automatic discovery of concepts from text. In: IEEE/WIC/ACM International Conference on Web Intelligence WI 2006, pp. 1046–1049 (December 2006)
20. Clack, C., Farringdon, J., Lidwell, P., Yu, T.: Autonomous document classification for business. In: International Conference on Autonomous Agents, pp. 201–208 (1997)
21. Cohen, W.: Learning to Classify English Text with ILP Methods. In: De Raedt, L. (ed.) Advances in Inductive Logic Programming, pp. 124–143. IOS Press, Amsterdam (1995)
22. Cohen, W., Hirsh, H.: Joins that generalize: Text classification using WHIRL. In: International ACM Conference on Knowledge Discovery and Data Mining - KDD 1998, pp. 169–173 (1998)
23. Cohen, W., Singer, Y.: Context-sensitive learning methods for text categorization. ACM Transactions on Information Systems 17(2), 141–173 (1999)
24. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine Learning 15(2), 201–221 (1994)
25. Cook, D., Holder, L. (eds.): Mining Graph Data. John Wiley & Sons, Chichester (2007)
26. Cooley, R.: Classification of news stories using support vector machines. In: International Joint Conference on Artificial Intelligence, Text Mining Workshop - IJCAI 1999 (1999)
27. Dagan, I., Engelson, S.: Committee-based sampling for training probabilistic classifiers. In: International Conference on Machine Learning - ICML 1995, pp. 150–157 (1995)
28. Dagan, I., Karov, Y., Roth, D.: Mistake driven learning in text categorization. In: Conference on Empirical Methods in Natural Language Processing - EMNLP 1997, pp. 55–63 (1997)
29. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic indexing. Journal of the American Society of Information Science 41(6), 391–407 (1990)
30. Drucker, H., Vapnik, V., Wu, D.: Automatic text categorization and its applications to text retrieval. IEEE Transactions on Neural Networks 10(5), 1048–1054 (1999)
31. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. John Wiley & Sons, Chichester (1993)
32. Dumais, S., Chen, H.: Hierarchical classification of Web content. In: Belkin, N.J., Ingwersen, P., Leong, M.-K. (eds.) International Conference on Research and Development in Information Retrieval - ACM SIGIR 2000, pp. 256–263 (2000)
33. Dumais, S., Platt, J., Heckerman, D., Sahami, M.: Inductive learning algorithms and representations for text categorization. In: International Conference on Information and Knowledge Management - ICIKM 1998, pp. 148–155. ACM Press, New York (1998)

34. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman & Hall, Boca Raton (1993)
35. European Commission. ICT - information and communication technologies - work programme 2007-2008 (2007)
36. Eyheramendy, S., Genkin, A., Ju, W., Lewis, D., Madigan, D.: Sparse bayesian classifiers for text categorization. Technical report, Department of Statistics, Rutgers University (2003)
37. Fawcett, T.: Roc graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Laboratories (2004)
38. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: International Conference on Machine Learning - ICML 1996, pp. 148–156 (1996)
39. Freund, Y., Seung, H., Shamir, E., Tishby, N.: Information, Prediction, and Query by Committee. In: Advances in Neural Information Processing Systems - NIPS 1993, pp. 483–490. Morgan Kaufmann Publishers Inc, San Francisco (1993)
40. Gale, W., Church, W., Yarowski, D.: A method for disambiguating word senses in a large corpus. Computers and the Humanities 26(5), 415–439 (1992)
41. Godbole, S.: Inter-class Relationships in Text Classification. PhD thesis, Indian Institute of Technology (2006)
42. Hayes, P., Anderson, P., Nirenburg, I., Schmandt, L.: TCS: A shell for content-based text categorization. In: IEEE Conference on Artificial Intelligence Applications, pp. 320–326. IEEE Press, Los Alamitos (1990)
43. Hayes, P., Weinstein, S.: CONSTRUE/TIS: A system for content-based indexing of a database of news stories. In: Conference on Innovative Applications of Artificial Intelligence - IAAI 1990, pp. 49–64 (1990)
44. Haykin, S.: Neural Networks: A Comprehensive Foundation. McGraw Hill, New York (2000)
45. Hotho, A., Nürnberger, A., Paass, G.: A brief survey of text mining. LDV Forum 20(1), 19–62 (2005)
46. Hu, X., Zhang, X., Lu, C., Park, E., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: Knowledge Discovery and Data Mining (KDD 2009), pp. 389–396 (2009)
47. Huang, J., Wang, G., Wang, Z.: Cross-subject page ranking based on text categorization. In: International Conference on Information and Automation, pp. 363–368 (2008)
48. Hull, D.: Improving text retrieval for the routing problem using latent semantic indexing. In: International Conference on Research and Development in Information Retrieval - ACM SIGIR 1994, pp. 282–289 (1994)
49. Jia, Z., Hu, M., Song, H., Hong, L.: Web text categorization for enterprise decision support based on SVMs: An application of GBODSS. In: ISNN 2009. LNCS, vol. 5552, pp. 753–762. Springer, Heidelberg (2009)
50. Jo, T., Yeom, G.: List based matching algorithm for classifying news articles in newspage.com. In: IEEE International Conference on System of Systems Engineering, pp. 1–5 (2008)
51. Joachims, T.: A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In: International Conference on Machine Learning - ICML 1997, pp. 143–151 (1997)
52. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
53. Joachims, T.: Transductive inference for text classification using support vector machines. In: International Conference on Machine Learning - ICML 1999, pp. 200–209. Morgan Kaufmann Publishers, San Francisco (1999)

54. Joachims, T.: Learning to Classify Text Using Support Vector Machines. Kluwer Academic Publishers, Dordrecht (2001)
55. Johansson, R., Nugues, P.: Sparse bayesian classification of predicate arguments. In: Conference on Computational Natural Language Learning - CoNLL 2005, pp. 177–180 (2005)
56. John, G., Kohavi, R., Peleger, K.: Irrelevant features and the subset selection problem. In: International Conference on Machine Learning - ICML 1994, pp. 121–129 (1994)
57. Khy, S., Ishikawa, Y., Kitagawa, H.: A novelty-based clustering method for on-line documents. World Wide Web 11(1), 1–37 (2008)
58. Kim, H., Howland, P., Park, H.: Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research (JLMR) 6, 37–53 (2005)
59. König, A., Brill, E.: Reducing the human overhead in text categorization. In: International ACM Conference on Knowledge Discovery and Data Mining - KDD 2006, pp. 598–603 (2006)
60. Kuncheva, L.: Combining Pattern Classifiers - Methods and Algorithms. Wiley, Chichester (2004)
61. Kuś, W.: Evolutionary optimization of forging anvils using grid based on alchemi framework. In: IEEE International Conference on e-Science and Grid Computing, pp. 121–125 (2006)
62. Kwok, J.: Automated text categorization using support vector machine. In: International Conference on Neural Information Processing - ICONI 1998, pp. 347–351 (1998)
63. Larkey, L.: A patent search and classification system. In: ACM Conference on Digital Libraries - DL 1999, pp. 179–187 (1999)
64. Larkey, L., Croft, W.: Combining classifiers in text categorization. In: International Conference on Research and Development in Information Retrieval - ACM SIGIR 1996, pp. 289–297 (1996)
65. Lee, C., Chiu, H., Yang, H.: A platform of biomedical literature mining for categorization of cancer related abstracts. In: Second International Conference on Innovative Computing, Information and Control, p. 174. IEEE Computer Society Press, Los Alamitos (2007)
66. Lewis, D.: Representation and Learning in Information Retrieval. PhD thesis, Computer Science Department; University of Massachusetts (1991)
67. Lewis, D.: An evaluation of phrasal and clustered representations on a text categorization. In: International Conference on Research and Development in Information Retrieval - ACM SIGIR 1992, pp. 37–50 (1992)
68. Lewis, D.: Feature selection and feature extraction for text categorization. In: Speech and Natural Language Workshop, pp. 212–217 (1992)
69. Lewis, D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: International Conference on Machine Learning - ICML 1994, pp. 148–156 (1994)
70. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: International Conference on Research and Development in InformationRetrieval - ACM SIGIR 1994, pp. 3–12 (1994)
71. Lewis, D., Ringuette, M.: A comparison of two learning algorithms for text categorization. In: Symposium on Document Analysis and Information Retrieval, pp. 81–93. University of Nevada (1994)
72. Lewis, D., Schapire, R., Callan, J., Papka, R.: Training algorithms for linear text classifiers. In: International Conference on Research and Development in Information Retrieval - ACM SIGIR 1996, pp. 298–306. ACM Press, New York (1996)
73. Lewis, D., Yang, Y., Rose, T., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research 5, 361–397 (2004)

74. Li, H., Yamanishi, K.: Text classification using esc-based stochastic decision lists. In: International ACM Conference on Information and Knowledge Management - CIKM 1999, pp. 122–130 (1999)
75. Li, Y., Jain, A.: Classification of text documents. Computing Journal 41(8), 537–546 (1998)
76. Liere, R., Tadepalli, P.: Active learning with committees for text categorization. In: Conference of the American Association for Artificial Intelligence - AAAI 1997, pp. 591–596 (1997)
77. Linstead, E., Rigor, P., Bajracharya, S., Lopes, C., Baldi, P.: Mining internet-scale software repositories. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems, vol. 20, pp. 929–936. MIT Press, Cambridge (2008)
78. Liu, B., Dai, Y., Li, X., Lee, W., Yu, P.: Building text classifiers using positive and unlabeled examples. In: IEEE International Conference on Data Mining - ICDM 2003, pp. 179–188. IEEE Computer Society, Los Alamitos (2003)
79. Liu, R.: Dynamic category profiling for text filtering and classification. Information Processing and Management: an International Journal 43(1), 154–168 (2007)
80. Lotrič, U., Silva, C., Ribeiro, B., Dobnikar, A.: Modeling execution times of data mining problems in grid environment. In: Trost, A., Zajc, B. (eds.) International ERK Conference, pp. 113–116. IEEE Press, Los Alamitos (2005)
81. MacKay, D.: The evidence framework applied to classification networks. Neural Computation 4(5), 720–736 (1992)
82. MacKay, D.: In: Domany, E., van Hemmen, J.L., Schulten, K. (eds.) Bayesian Methods for Backpropagation Networks, Models of Neural Networks III, ch. 6, pp. 211–254. Springer, Newyork (1994)
83. Makrehchi, M., Kamel, M.: A text classification framework with a local feature ranking for learning socialnetworks. In: Seventh IEEE Conference on Data Mining, pp. 589–594. IEEE Computer Society, Los Alamitos (2007)
84. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
85. Masand, B.: Optimising Confidence of Text Classification By Evolution of Symbolic Expressions. In: Advances in Genetic Programming, ch. 21, pp. 459–476. MIT Press, Cambridge (1994)
86. McCallum, A., Nigam, K.: Employing EM and pool-based active learning for text classification. In: International Conference on Machine Learning - ICML 1998, pp. 350–358. Morgan Kaufmann Publishers, San Francisco (1998)
87. Mei, J., Zhang, W., Wang, S.: Grid enabled problem solving environments for text categorization. In: IEEE International Conference on e-Science and Grid Computing, pp. 106–110 (2006)
88. Melab, N., Cahon, S., Talbi, E.: Grid computing for parallel bioinspired algorithms. Journal of Parallel and Distributed Computing 66(9), 1052–1061 (2006)
89. Minier, Z., Bodo, Z., Csato, L.: Wikipedia-based kernels for text categorization. In: International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, pp. 157–164. SYNASC (2007)
90. Mitchell, T.: Machine Learning. McGraw Hill, New York (1996)
91. Mladenić, D.: Feature subset selection in text learning. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 95–100. Springer, Heidelberg (1998)

92. Moulinier, I., Ganascia, J.: Applying an Existing Machine Learning Algorithm to Text Categorization. In: Wermter, S., Riloff, E., Schaler, G. (eds.) Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing, pp. 343–354. Springer, Heidelberg (1996)
93. Moulinier, I., Raskinis, G., Ganascia, J.: Text categorization: a symbolic approach. In: Annual Symposium on Document Analysis and Information Retrieval - SDAIR 1996, pp. 87–99 (1996)
94. Nedjah, N., Mourelle, L., Kacprzyk, J., França, F., Souza, A. (eds.): Intelligent Text Categorization and Clustering. Studies in Computational Intelligence, vol. 164. Springer, Heidelberg (2009)
95. Ng, A., Jordan, M.: On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regressionand Naïve Bayes. In: Dietterich, T., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems - NIPS 2002, vol. 14, pp. 609–616. MIT Press, Cambridge (2002)
96. Ng, H., Goh, W., Low, K.: Feature selection, perceptron learning, and a usability case study for text categorization. In: International Conference on Research and Development in InformationRetrieval - ACM SIGIR 1997, pp. 67–73 (1997)
97. Nigam, K., Mccallum, A., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Machine Learning 39(2/3), 103–134 (2000)
98. Pal, S.K., Narayan, B.L.: A web surfer model incorporating topic continuity. IEEE Transactions on Knowledge and Data Engineering 17(5), 726–729 (2005)
99. Park, C.: Dimension reduction using least squares regression in multi-labeled text categorization. In: IEEE International Conference on Computer and Information Technology, pp. 71–76 (2008)
100. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 6(3), 21–45 (2006)
101. Porter, M.: An agorithm for suffix stripping. Program 14(3), 130–137 (1980)
102. Potamias, G., Koumakis, L., Moustakis, V.: Enhancing web based services by coupling document classification with user profile. In: The International Conference on Computer as a Tool - EUROCON 2005, vol. 1, pp. 205–208 (2005)
103. Quinn, M.: Parallel Programming in C with MPI and OpenMP. McGraw-Hill, New York (2003)
104. Ribeiro, B., Silva, C., Vieira, A., Neves, J.: Extracting discriminative features using non-negative matrix factorization in financial distress data. In: ICANNGA 2009. LNCS, Springer, Heidelberg (2009)
105. Rigutini, L., Di Iorio, E., Ernandes, M., Maggini, M.: Semantic labeling of data by using the web. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, December 2006, pp. 638–641 (2006)
106. Rocchio, J.: Relevance Feedback in Information Retrieval. In: Salton, G. (ed.) The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, Englewood Cliffs (1971)
107. Rose, T., Stevenson, M., Whitehead, M.: The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In: International Conference on Language Resources and Evaluation - LREC 2002, pp. 29–31 (2002)
108. Ruiz, M., Srinivasan, P.: Automatic text categorization and its application to text retrieval. IEEE Transactions on Knowledge and Data Engineering 11(6), 865–879 (1999)
109. Ruiz, M., Srinivasan, P.: Hierarchical text categorization using neural networks. Information Retrieval 5(1), 87–118 (2002)
110. Saldarriaga, S., Morin, E., Viard-Gaudin, C.: Categorization of on-line handwritten documents. In: International Workshop on Document Analysis Systems, pp. 95–102 (2008)

111. Sarnovský, M., Butka, P.: Grid-enabled support for classification and clustering of textual documents. In: Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics, pp. 265–275 (2007)
112. Schaffer, C.: A conservation law for generalization performance. In: International Conference on Machine Learning - ICML 1994, pp. 259–265 (1994)
113. Schapire, R., Singer, Y.: Boostexter: A boosting-based system for text categorization. Machine Learning 39(2/3), 135–168 (2000)
114. Schapire, R., Singer, Y., Singhal, A.: Boosting and rocchio applied to text filtering. In: International Conference on Research and Development in Information Retrieval - ACM SIGIR 1998, pp. 215–223 (1998)
115. Schölkopf, B., Burges, C., Smola, A.: Advances in Kernel methods, pp. 1–15. MIT Press, Cambridge (1999)
116. Schohn, G., Cohn, D.: Less is more: Active learning with support vector machines. In: International Conference on Machine Learning - ICML 2000, pp. 839–846. Morgan Kaufmann, San Francisco (2000)
117. Schölkopf, B.: Support Vector Learning. R. Oldenbourg Verlag (1997)
118. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge (2002)
119. Schölkopf, B., Smola, A., Müller, K.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10(5), 1299–1319 (1998)
120. Schütze, H., Hull, D., Pendersen, J.: A comparison of classifiers and document representations for the routing problem. In: International Conference on Research and Development in InformationRetrieval - ACM SIGIR 1995, pp. 229–237 (1995)
121. Sebastiani, F.: A tutorial on automated text categorisation. In: Argentinian Symposium on Artificial Intelligence - ASAI 1999, pp. 7–35 (1999)
122. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
123. Sebastiani, F.: Classification of text, automatic. In: Brown, K. (ed.) The Encyclopedia of Language and Linguistics, 2nd edn., vol. 14, pp. 457–462. Elsevier Science Publishers, Amsterdam (2006)
124. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
125. Silva, C., Ribeiro, B.: Rare class text categorization with SVM ensemble. Journal of Electrotechnical Review (Przeglad Elektrotechniczny) 1, 28–31 (2006)
126. Silva, C., Ribeiro, B.: On text-based mining with active learning and background knowledge using SVM. Journal of Soft Computing - A Fusion of Foundations, Methodologies and Applications 11(6), 519–530 (2007)
127. Silva, C., Ribeiro, B.: The importance of stop word removal on recall values in text categorization. In: IEEE International Joint Conference on Neural Networks - IJCNN 2003, pp. 1661–1666 (2003)
128. Silva, C., Ribeiro, B.: Labeled and unlabeled data in text categorization. In: IEEE International Joint Conference on Neural Networks - IJCNN 2004, pp. 2971–2976 (2004)
129. Silva, C., Ribeiro, B.: Margin-based active learning and background knowledge in text mining. In: International Conference on Hybrid Intelligent Systems - HIS 2004, pp. 8–13 (2004)
130. Silva, C., Ribeiro, B.: Text classification from partially labeled distributed data. In: International Conference on Adaptive and Natural Computing Algorithms - ICANNGA 2005. LNCS, pp. 445–448. Springer, Heidelberg (2005)
131. Silva, C., Ribeiro, B.: Automated learning of RVM for large scale text sets: Divide to conquer. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 878–886. Springer, Heidelberg (2006)

132. Silva, C., Ribeiro, B.: Ensemble RVM for text classification. In: International Conference on Neural Information Processing - ICONIP2006 (2006)
133. Silva, C., Ribeiro, B.: Fast-decision SVM ensemble text classifier using cluster computing. In: International Conference on Neural, Parallel & Scientific Computations - ICNPSC 2006, pp. 253–259 (2006)
134. Silva, C., Ribeiro, B.: Two-level hierarchical hybrid SVM-RVM classification model. In: IEEE International Conference on Machine Learning and Applications - ICMLA 2006, pp. 89–94 (2006)
135. Silva, C., Ribeiro, B.: Combining active learning and relevance vector machines for text classification. In: IEEE International Conference on Machine Learning and Applications - ICMLA 2007, pp. 130–135 (2007)
136. Silva, C., Ribeiro, B.: RVM ensemble for text classification. International Journal of Computational Intelligence Research 3(1), 31–35 (2007)
137. Silva, C., Ribeiro, B.: Towards expanding relevance vector machines to large scale datasets. International Journal of Neural Systems 18(1), 45–58 (2008)
138. Silva, C., Ribeiro, B.: Improving text classification performance with incremental background knowledge. In: ICANN 2009, Part I. LNCS, vol. 5768. Springer, Heidelberg (2009)
139. Silva, C., Ribeiro, B.: Improving visualization, scalability and performance of multiclass problems with SVM manifold learning. In: ICANNGA 2009. LNCS. Springer, Heidelberg (2009)
140. Silva, C., Ribeiro, B., Lotrič, U.: Speeding-up text classification in a grid computing environment. In: IEEE International Conference on Machine Learning and Applications - ICMLA 2005, pp. 113–116 (2005)
141. Silva, C., Ribeiro, B., Lotrič, U., Dobnikar, A.: Distributed ensemble learning in text classification. In: International Conference on Enterprise Information Systems - ICEIS 2008, pp. 420–423 (2008)
142. Silva, C., Ribeiro, B., Sung, A.: Boosting RVM classifiers for large data sets. In: Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B. (eds.) ICANNGA 2007. LNCS, vol. 4432, pp. 228–237. Springer, Heidelberg (2007)
143. Song, Y., Zhang, L., Giles, C.: Automatic tag recommendation algorithms for social recommender systems. ACM Transactions on the Web, TWEB (2009)
144. Szummer, M.: Learning from Partially Labeled Data. PhD thesis, Massachusetts Institute of Technology (2002)
145. Tenenbaum, J., Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
146. Tian, Y., Yang, Q., Huang, T., Lin, C., Gao, W.: Learning contextual dependency network models for link-based classification. IEEE Transactions on Knowledge and Data Engineering 18(11), 1482–1496 (2006)
147. Tipping, M.: Sparse bayesian learning and the relevance vector machine. Journal of Machine Learning Research 1, 211–214 (2001)
148. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of Machine Learning Research 2, 45–66 (2001)
149. Tzeras, K., Hartmann, S.: Automatic indexing based on bayesian inference networks. In: International Conference on Research and Development in Information Retrieval - ACM SIGIR 1993, pp. 22–34 (1993)
150. van Rijsbergen, C.: Information Retrieval. Butterworths Ed. (1979)
151. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1995)
152. Vapnik, V.: Statistical Learning Theory. Wiley, Chichester (1998)

153. Vert, J., Matsui, T., Satoh, S., Uchiyama, Y.: High-level feature extraction using SVM with walk-based graph kernel. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009 (2009)
154. Vert, J., Yamanishi, Y.: Supervised graph inference. Advances in Neural Information Processing Systems 17, 1433–1440 (2005)
155. Wagner, J., Foster, J., Genabith, J.: A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - EMNLP-CoNLL 2007, pp. 112–121 (2007)
156. Wang, T., Desai, B.: An approach for text categorization in digital library. In: 11th International Database Engineering and Applications Symposium, pp. 21–27 (2007)
157. Wei, L., Yang, Y., Nishikawa, R., Wernick, M., Edwards, A.: Relevance vector machine for automatic detection of clustered microcalcifications. IEEE Transactions on Medical Imaging 24(10), 1278–1285 (2005)
158. Weiss, S., Apté, C., Damerau, F., Johnson, D., Oles, F., Goetz, T., Hampp, T.: Maximizing text-mining performance. IEEE Intelligent Systems 14(4), 63–69 (1999)
159. Wiener, E., Pedersen, J., Weigend, A.: A neural network approach to topic spotting. In: Annual Symposium on Document Analysis and Information Retrieval - SDAIR 1995, pp. 317–332 (1995)
160. Wolpert, D., Macready, W.: No free lunch theorems for search. Technical Report SFI-TR-95-02-010, Santa Fe Institute (1995)
161. Wolpert, D., Macready, W.: Coevolutionary free lunches. IEEE Transactions on Evolutionary Computation 9(6), 721–735 (2005)
162. Wu, Z., Hsu, L., Tan, C.: A survey of statistical approaches to natural language processing. Technical Report TRA4/92, Department of Information Systems and Computer Science, National University of Singapore (1992)
163. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: International Conference on Research and Development in Information Retrieval - ACM SIGIR 1999, pp. 42–49 (1999)
164. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: International Conference on Machine Learning - ICML 1997, pp. 412–420 (1997)
165. Yu, L., Wang, S., Lai, K., Wu, Y.: A framework of web-based text mining on the grid. In: IEEE International Conference on Next Generation Web Services Practices, pp. 97–102 (2005)
166. Zhang, T., Oles, F.: A probability analysis on the value of unlabeled data for classification problems. In: International Conference on Machine Learning - ICML 2000, pp. 1191–1198 (2000)
167. Zhang, X., Mei, J., Wang, S., Zhang, W.: Web services enabled text categorization system: Service infrastructure building. International Journal of Computer Science and Network Security 7(2), 73–77 (2007)
168. Zheng, Y., Duan, L., Tian, Q., Jin, J.: Tv commercial classification by using multi-modal textual information. In: IEEE International Conference on Multimedia and Expo, pp. 497–500 (2006)
169. Zhou, D., Huang, J., Schölkopf, B.: Learning from labeled and unlabeled data on a directed graph. In: International Conference on Machine Learning - ICML -2005, pp. 1036–1043 (2005)
170. Zhou, G., Zhang, J., Su, J., Shen, D., Tan, C.: Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 20(7), 1178–1190 (2004)

Index