Email Classification Using Co-EM

Email Classification Using Co-EM Project Report

Aravind Sankar (CS11B033)
Sriram V (CS11B058)

Table of Contents

1 Introduction

The broad area of Text Classification has been a topic of research for a very long time now, and has been well researched especially in the case of supervised learning of classifiers. The traditional problems solved in the past by text classification include automatically cataloging news articles, web pages and spam detection in emails, etc. Text classification as defined by [1], is the activity of labeling natural language texts with thematic categories from a predefined set. Most of the earlier approaches assume the existence of sufficient training data, which has labeled instances of documents with their corresponding labels. As the process of labeling here requires human effort, our focus is therefore directed towards tackling the problem of classification of limited labeled data, otherwise referred to as semi-supervised learning, assuming we have a large number of unlabeled documents.

Although text classification is a problem that has been looked into right from the 1960s, email classification is highly relevant in today's fast-paced world, where one does not have the time to sift through vast amounts of mail but would rather like information clustered together, so that one may disregard the irrelevant ones such as promotional offers, and focus attention on the important ones at hand. Triaging has gained a lot of focus in recent times, with GMail launching personalized mail classification, and with other services such as Mailbox and Boxer focusing on improving user efficiency. Thus, this problem is of immense interest in recent times, and improving performance and accuracy, without existence of enough labeled email, is of prime concern.

One of the most popular problems addressed by email classification is that of spam filtering, i.e. separating emails into spam and ham. This problem has been widely addressed, with various classifiers and feature selection and extraction methods experimented with. Prior work has shown great results on the Enron corpus, which is the most popular dataset used for this problem. We wish to focus on the problem of categorizing emails into categories, which would help a user identify the emails important to him. If we have a idea about the areas of interest of a user, then the emails could be easily filtered according to the category and the relevant mails can be displayed first, hence making it personalized to the user.

Before we describe the exact problem we solve and our approach, a brief survey of existing techniques to solve email classification is shown in the following section.

2 Literature Survey

As mentioned earlier, spam filtering was one of the popular problems addressed. Initially, the mails were classified as spam using the presence of certain predefined tokens in the email body. As these methods require manual specification of these tokens, the focus shifted to learning, which required the mails to defined as a set of features. Among the various feature extraction techniques, the Bag-of-Words feature representation has been widely used. This representation gives a set of binary features, for each mail, where each feature shows the presence or absence of a particular term. Some of the other methods such as [2] include other spam features, that are domain specific. Other techniques such as [3] have used temporal features which identify temporal relations in an email sequence in the form of temporal sequential patterns using the timestamps. Behavioral features have also been identified by [4], by looking at outgoing messages from a user.

Once the features have been extracted from the emails, the standard supervised classification techniques have been put to use. The most predominant ones, being the Naive Bayes and the SVM classifiers. The Naive Bayes classifier assumes that each of the features $f_1, f_2, ...f_n$ of the email are conditionally independent given the class label C. Surprisingly, this has proved to quite effective for email classification. Studies have been performed, comparing different classification approaches, eg: the Naive Bayes has been compared with Rule learning approaches by [5] and has been shown that Naive Bayes performs much better. SVM's have also been widely used, which try to separate the two classes by an optimum hyperplane. Another comparative study by [6] showed the effectiveness of a decision tree classifier in spam filtering in comparison to SVM's and Neural Networks.

The unsupervised approaches for email classification are broadly based upon 2 techniques, namely EM (Expectation Maximization) and Co-Training. These approaches typically start with a very small set of labeled instances and slowly label each of the unlabeled instances in subsequent iterations. The EM algorithm has been traditionally used to solve incomplete data problems, in many different areas. Here, the unlabeled emails constitute the incomplete data of the problem. The EM algorithm for text classification first introduced by [7] first trains a classifier with only the available labeled documents, and then uses the classifier to assign probabilistically-weighted class labels to each unlabeled document by calculating the expectation of the missing class labels. The idea of co-training, on the other hand is different. Co-training was used in email classification first by [8]. It starts with just a few labeled emails, and builds an initial weak classifier. It assumes the existence of 2 sets of redundant features but sufficient on their own for correct classification. We can see that in case of email, we can think of the email body and email header (including subject) as a set of 2 independent feature sets. [8] have shown that SVM's have performed better than the Naive Bayes classifier when Co-training is used.

Our approach draws inspiration from Co-Training and EM, and uses the Co-EM approach, which was proposed by [9]. To the best of our knowledge, Co-EM has not been experimented with, for the purpose of email classification. We use Co-EM for email classification, and compare with Co-Training and evaluate it's effectiveness.

3 Methodology

Before going into the details of our approach, we provide a brief description of the Co-Training algorithm, as applied to email classification by [8].

Co-Training :

Given 2 redundantly sufficient sets of features F_1 and F_2, L - the set of labeled mails (typically small) and U - the set of unlabeled mails, the algorithm switches between 2 classifiers and add mails to L, in each iteration.

```
Data: $L$,$U$,$k$ (parameter)
Result: Fully labeled set $L$
while Any documents are present in $U$ do
    Learn classifier $C_1$ from $F_1$ on $L$;
    Learn classifier $C_1$ from $F_2$ on $L$;
    for Classifier $C$ in $\{C_1, C_2\}$ do
        $C$ labels mails in $U$ based on it's feature set.
        Choose $k$ most confidently predicted mails from each class, in $U$.
        Remove this set of mails from $U$ and add them to $L$, with the
        labels predicted by $C$.
    end
end
```

Algorithm 1: Co-Training

Here, the set F_1 of features used are extracted from the body of the mail, and the set F_2 from the mail header, which includes information about the sender, receiver and the subject of the mail. As these two pieces of information are mostly disjoint, and as subject and sender may prove to be a good discriminator in classifying emails, we choose this split.

4 Experiments and Results

5 Conclusions and Future Work

5.1 How to Invoke the LLNCS Document Class

The LLNCS class is an extension of the standard LaTeX "article" document class. Therefore you may use all "article" commands for the body of your contribution to prepare your manuscript. LLNCS class is invoked by replacing "article" by "llncs" in the first line of your document:

```
\documentclass{llncs}
%
\begin{document}
  <Your contribution>
\end{document}
```

5.2 Contributions Already Coded with LaTeX without the LLNCS document class

If your file is already coded with LaTeX you can easily adapt it a posteriori to the LLNCS document class.

Please refrain from using any LaTeX or TeX commands that affect the layout or formatting of your document (i.e. commands like `\textheight`, `\vspace`, `\headsep` etc.). There may nevertheless be exceptional occasions on which to use some of them.

The LLNCS document class has been carefully designed to produce the right layout from your LaTeX input. If there is anything specific you would like to do and for which the style file does not provide a command, *please contact us*. Same holds for any error and bug you discover (there is however no reward for this – sorry).

6 General Rules for Coding Formulas

With mathematical formulas you may proceed as described in Sect. 3.3 of the *LaTeX User's Guide & Reference Manual* by Leslie Lamport (2nd ed. 1994), Addison-Wesley Publishing Company, Inc.

Equations are automatically numbered sequentially throughout your contribution using arabic numerals in parentheses on the right-hand side.

When you are working in math mode everything is typeset in italics. Sometimes you need to insert non-mathematical elements (e.g. words or phrases). Such insertions should be coded in roman (with `\mbox`) as illustrated in the following example:

Sample Input

```
\begin{equation}
  \left(\frac{a^{2} + b^{2}}{c^{3}} \right) = 1 \quad
  \mbox{ if } c\neq 0 \mbox{ and if } a,b,c\in \bbbr \enspace .
\end{equation}
```

Sample Output

$$\left(\frac{a^{2} + b^{2}}{c^{3}} \right) = 1 \quad \mbox{ if } c\neq 0 \mbox{ and if } a,b,c\in \mathbb{R} \enspace . \tag{1}$$

If you wish to start a new paragraph immediately after a displayed equation, insert a blank line so as to produce the required indentation. If there is no new paragraph either do not insert a blank line or code `\noindent` immediately before continuing the text.

Please punctuate a displayed equation in the same way as other ordinary text but with an `\enspace` before end punctuation.

Note that the sizes of the parentheses or other delimiter symbols used in equations should ideally match the height of the formulas being enclosed. This is automatically taken care of by the following LaTeX commands:

`\left(` or `\left[` and `\right)` or `\right]`.

6.1 Italic and Roman Type in Math Mode

a) In math mode LaTeX treats all letters as though they were mathematical or physical variables, hence they are typeset as characters of their own in italics. However, for certain components of formulas, like short texts, this would be incorrect and therefore coding in roman is required. Roman should also be used for subscripts and superscripts *in formulas* where these are merely labels and not in themselves variables, e.g. T_{eff} *not* T_{eff}, T_{K} *not* T_K (K = Kelvin), m_{e} *not* m_e (e = electron). However, do not code for roman if the sub/superscripts represent variables, e.g. $\sum_{i=1}^{n} a_i$.

b) Please ensure that *physical units* (e.g. pc, erg s^{-1} K, cm^{-3}, W m^{-2} Hz^{-1}, m kg s^{-2} A^{-2}) and *abbreviations* such as Ord, Var, GL, SL, sgn, const. are always set in roman type. To ensure this use the `\mathrm` command: `\mathrm{Hz}`. On p. 44 of the *LaTeX User's Guide & Reference Manual* by Leslie Lamport you will find the names of common mathematical functions, such as log, sin, exp, max and sup. These should be coded as `\log`, `\sin`, `\exp`, `\max`, `\sup` and will appear in roman automatically.

c) Chemical symbols and formulas should be coded for roman, e.g. Fe not *Fe*, $\mathrm{H_2O}$ not H_2O.

d) Familiar foreign words and phrases, e.g. et al., a priori, in situ, bremsstrahlung, eigenvalues should not be italicized.

7 How to Edit Your Input (Source) File

7.1 Headings

All words in headings should be capitalized except for conjunctions, prepositions (e.g. on, of, by, and, or, but, from, with, without, under) and definite and indefinite articles (the, a, an) unless they appear at the beginning. Formula letters must be typeset as in the text.

7.2 Capitalization and Non-capitalization

a) The following should always be capitalized:
   - Headings (see preceding Sect. 7.1)
   - Abbreviations and expressions in the text such as Fig(s)., Table(s), Sect(s)., Chap(s)., Theorem, Corollary, Definition etc. when used with numbers, e.g. Fig. 3, Table 1, Theorem 2.

   Please follow the special rules in Sect. 7.3 for referring to equations.

b) The following should *not* be capitalized:
   - The words figure(s), table(s), equation(s), theorem(s) in the text when used without an accompanying number.
   - Figure legends and table captions except for names and abbreviations.

7.3 Abbreviation of Words

a) The following *should* be abbreviated when they appear in running text *unless* they come at the beginning of a sentence: Chap., Sect., Fig.; e.g. The results are depicted in Fig. 5. Figure 9 reveals that . . . .

   Please note: Equations should usually be referred to solely by their number in parentheses: e.g. (14). However, when the reference comes at the beginning of a sentence, the unabbreviated word “Equation” should be used: e.g. Equation (14) is very important. However, (15) makes it clear that . . . .

b) If abbreviations of names or concepts are used throughout the text, they should be defined at first occurrence, e.g. Plurisubharmonic (PSH) Functions, Strong Optimization (SOPT) Problem.

8 How to Code the Beginning of Your Contribution

The title of a single contribution (it is mandatory) should be coded as follows:

```
\title{<Your contribution title>}
```

All words in titles should be capitalized except for conjunctions, prepositions (e.g. on, of, by, and, or, but, from, with, without, under) and definite and indefinite articles (the, a, an) unless they appear at the beginning. Formula letters must be typeset as in the text. Titles have no end punctuation.

If a long `\title` must be divided please use the code `\\` (for new line).

If you are to produce running heads for a specific volume the standard (of no such running heads) is overwritten with the `[runningheads]` option in the `\documentclass` line. For long titles that do not fit in the single line of the running head a warning is generated. You can specify an abbreviated title for the running head on odd pages with the command

```
\titlerunning{<Your abbreviated contribution title>}
```

There is also a possibility to change the text of the title that goes into the table of contents (that's for volume editors only – there is no table of contents for a single contribution). For this use the command

```
\toctitle{<Your changed title for the table of contents>}
```

An optional subtitle may follow then:

```
\subtitle{<subtitle of your contribution>}
```

Now the name(s) of the author(s) must be given:

```
\author{<author(s) name(s)>}
```

Numbers referring to different addresses or affiliations are to be attached to each author with the `\inst{<no>}` command. If there is more than one author, the order is up to you; the `\and` command provides for the separation.

If you have done this correctly, this entry now reads, for example:

```
\author{Ivar Ekeland\inst{1} \and Roger Temam\inst{2}}
```

The first name[1] is followed by the surname.

As for the title there exist two additional commands (again for volume editors only) for a different author list. One for the running head (on odd pages) – if there is any:

```
\authorrunning{<abbreviated author list>}
```

And one for the table of contents where the affiliation of each author is simply added in braces.

```
\tocauthor{<enhanced author list for the table of contents>}
```

Next the address(es) of institute(s), company etc. is (are) required. If there is more than one address, the entries are numbered automatically with `\and`, in the order in which you type them. Please make sure that the numbers match those placed next to to the authors' names to reflect the affiliation.

```
\institute{<name of an institute>
\and <name of the next institute>
\and <name of the next institute>}
```

[1] Other initials are optional and may be inserted if this is the usual way of writing your name, e.g. Alfred J. Holmes, E. Henry Green.

In addition, you can use

```
\email{<email address>}
```

to provide your email address within `\institute`. If you need to typeset the tilde character – e.g. for your web page in your unix system's home directory – the `\homedir` command will happily do this. Please note that, if your email address is given in your paper, it will also be included in the meta data of the online version.

If footnote like things are needed anywhere in the contribution heading please code (immediately after the word where the footnote indicator should be placed):

```
\thanks{<text>}
```

`\thanks` may only appear in `\title`, `\author` and `\institute` to footnote anything. If there are two or more footnotes or affiliation marks to a specific item separate them with `\fnmsep` (i.e. *f*oot*n*ote *m*ark *sep*arator).

The command

```
\maketitle
```

then formats the complete heading of your article. If you leave it out the work done so far will produce *no* text.

Then the abstract should follow. Simply code

```
\begin{abstract}
<Text of the summary of your article>
\end{abstract}
```

or refer to the demonstration file `llncs.dem` for an example or to the *Sample Input* on p. 14.

Remark to Running Heads and the Table of Contents

If you are the author of a single contribution you normally have no running heads and no table of contents. Both are done only by the editor of the volume or at the printers.

9 Special Commands for the Volume Editor

The volume editor can produce a complete camera ready output including running heads, a table of contents, preliminary text (frontmatter), and index or glossary. For activating the running heads there is the class option `[runningheads]`.

The table of contents of the volume is printed wherever `\tableofcontents` is placed. A simple compilation of all contributions (fields `\title` and `\author`) is done. If you wish to change this automatically produced list use the commands

```
\titlerunning  \toctitle
\authorrunning \tocauthor
```

to enhance the information in the specific contributions. See the demonstration file `llncs.dem` for examples.

An additional structure can be added to the table of contents with the `\addtocmark{<text>}` command. It has an optional numerical argument, a digit from 1 through 3. 3 (the default) makes an unnumbered chapter like entry in the table of contents. If you code `\addtocmark[2]{text}` the corresponding page number is listed also, `\addtocmark[1]{text}` even introduces a chapter number beyond it.

10 How to Code Your Text

The contribution title and all headings should be capitalized except for conjunctions, prepositions (e.g. on, of, by, and, or, but, from, with, without, under) and definite and indefinite articles (the, a, an) unless they appear at the beginning. Formula letters must be typeset as in the text.

Headings will be automatically numbered by the following codes.

Sample Input

```
\section{This is a First-Order Title}
\subsection{This is a Second-Order Title}
\subsubsection{This is a Third-Order Title.}
\paragraph{This is a Fourth-Order Title.}
```

`\section` and `\subsection` have no end punctuation.
`\subsubsection` and `\paragraph` need to be punctuated at the end.

In addition to the above-mentioned headings your text may be structured by subsections indicated by run-in headings (theorem-like environments). All the theorem-like environments are numbered automatically throughout the sections of your document – each with its own counter. If you want the theorem-like environments to use the same counter just specify the documentclass option `envcountsame`:

```
\documentclass[envcountsame]{llncs}
```

If your first call for a theorem-like environment then is e.g. `\begin{lemma}`, it will be numbered 1; if corollary follows, this will be numbered 2; if you then call lemma again, this will be numbered 3.

But in case you want to reset such counters to 1 in each section, please specify the documentclass option `envcountreset`:

```
\documentclass[envcountreset]{llncs}
```

Even a numbering on section level (including the section counter) is possible with the documentclass option `envcountsect`.

11 Predefined Theorem like Environments

The following variety of run-in headings are at your disposal:

a) **Bold** run-in headings with italicized text as built-in environments:

```
\begin{corollary} <text> \end{corollary}
\begin{lemma} <text> \end{lemma}
\begin{proposition} <text> \end{proposition}
\begin{theorem} <text> \end{theorem}
```

b) The following generally appears as *italic* run-in heading:

```
\begin{proof} <text>    \qed    \end{proof}
```

It is unnumbered and may contain an eye catching square (call for that with `\qed`) before the environment ends.

c) Further *italic* or **bold** run-in headings with roman environment body may also occur:

```
\begin{definition} <text> \end{definition}
\begin{example} <text> \end{example}
\begin{exercise} <text> \end{exercise}
\begin{note} <text> \end{note}
\begin{problem} <text> \end{problem}
\begin{question} <text> \end{question}
\begin{remark} <text> \end{remark}
\begin{solution} <text> \end{solution}
```

12 Defining your Own Theorem like Environments

We have enhanced the standard `\newtheorem` command and slightly changed its syntax to get two new commands `\spnewtheorem` and `\spnewtheorem*` that now can be used to define additional environments. They require two additional arguments namely the type style in which the keyword of the environment appears and second the style for the text of your new environment.

`\spnewtheorem` can be used in two ways.

12.1 Method 1 *(preferred)*

You may want to create an environment that shares its counter with another environment, say *main theorem* to be numbered like the predefined *theorem*. In this case, use the syntax

```
\spnewtheorem{<env_nam>}[<num_like>]{<caption>}
{<cap_font>}{<body_font>}
```

Here the environment with which the new environment should share its counter is specified with the optional argument `[<num_like>]`.

Sample Input

```
\spnewtheorem{mainth}[theorem]{Main Theorem}{\bfseries}{\itshape}
\begin{theorem} The early bird gets the worm. \end{theorem}
\begin{mainth} The early worm gets eaten. \end{mainth}
```

Sample Output

Theorem 3. *The early bird gets the worm.*

Main Theorem 4. The early worm gets eaten.

The sharing of the default counter (`[theorem]`) is desired. If you omit the optional second argument of `\spnewtheorem` a separate counter for your new environment is used throughout your document.

12.2 Method 2 *(assumes* `[envcountsect]` *documentstyle option)*

```
\spnewtheorem{<env_nam>}{<caption>}[<within>]
{<cap_font>}{<body_font>}
```

This defines a new environment `<env_nam>` which prints the caption `<caption>` in the font `<cap_font>` and the text itself in the font `<body_font>`. The environment is numbered beginning anew with every new sectioning element you specify with the optional parameter `<within>`.

Example

```
\spnewtheorem{joke}{Joke}[subsection]{\bfseries}{\rmfamily}
```

defines a new environment called `joke` which prints the caption **Joke** in boldface and the text in roman. The jokes are numbered starting from 1 at the beginning of every subsection with the number of the subsection preceding the number of the joke e.g. 7.2.1 for the first joke in subsection 7.2.

12.3 Unnumbered Environments

If you wish to have an unnumbered environment, please use the syntax

```
\spnewtheorem*{<env_nam>}{<caption>}{<cap_font>}{<body_font>}
```

13 Program Codes

In case you want to show pieces of program code, just use the `verbatim` environment or the `verbatim` package of LaTeX. (There also exist various pretty printers for some programming languages.)

Sample Input (of a simple contribution)

```
\title{Hamiltonian Mechanics}

\author{Ivar Ekeland\inst{1} \and Roger Temam\inst{2}}

\institute{Princeton University, Princeton NJ 08544, USA
\and
Universit\'{e} de Paris-Sud,
Laboratoire d'Analyse Num\'{e}rique, B\^{a}timent 425,\\
F-91405 Orsay Cedex, France}

\maketitle
%
\begin{abstract}
This paragraph shall summarize the contents of the paper
in short terms.
\end{abstract}
%
\section{Fixed-Period Problems: The Sublinear Case}
%
With this chapter, the preliminaries are over, and we begin the
search for periodic solutions \dots
%
\subsection{Autonomous Systems}
%
In this section we will consider the case when the Hamiltonian
$H(x)$ \dots
%
\subsubsection*{The General Case: Nontriviality.}
%
We assume that $H$ is
$\left(A_{\infty}, B_{\infty}\right)$-subqua\-dra\-tic
at infinity, for some constant \dots
%
\paragraph{Notes and Comments.}
The first results on subharmonics were \dots
%
\begin{proposition}
Assume $H'(0)=0$ and $ H(0)=0$. Set \dots
\end{proposition}
\begin{proof}[of proposition]
Condition (8) means that, for every $\delta'>\delta$, there is
some $\varepsilon>0$ such that \dots \qed
\end{proof}
%
```

```
\begin{example}[\rmfamily (External forcing)]
Consider the system \dots
\end{example}
\begin{corollary}
Assume $H$ is $C^{2}$ and
$\left(a_{\infty}, b_{\infty}\right)$-subquadratic
at infinity. Let \dots
\end{corollary}
\begin{lemma}
Assume that $H$ is $C^{2}$ on $\bbbr^{2n}\backslash \{0\}$
and that $H''(x)$ is \dots
\end{lemma}
\begin{theorem}[(Ghoussoub-Preiss)]
Let $X$ be a Banach Space and $\Phi:X\to\bbbr$ \dots
\end{theorem}
\begin{definition}
We shall say that a $C^{1}$ function $\Phi:X\to\bbbr$
satisfies \dots
\end{definition}
```

Sample Output (follows on the next page together with examples of the above run-in headings)

Hamiltonian Mechanics

Ivar Ekeland[1] and Roger Temam[2]

[1] Princeton University, Princeton NJ 08544, USA
[2] Université de Paris-Sud, Laboratoire d'Analyse Numérique, Bâtiment 425,
F-91405 Orsay Cedex, France

Abstract. This paragraph shall summarize the contents of the paper in short terms.

1 Fixed-Period Problems: The Sublinear Case

With this chapter, the preliminaries are over, and we begin the search for periodic solutions . . .

1.1 Autonomous Systems

In this section we will consider the case when the Hamiltonian $H(x)$ . . .

The General Case: Nontriviality. We assume that H is (A_∞, B_∞)-subquadratic at infinity, for some constant . . .

Notes and Comments. The first results on subharmonics were . . .

Proposition 1. *Assume $H'(0) = 0$ and $H(0) = 0$. Set . . .*

Proof (of proposition). Condition (8) means that, for every $\delta' > \delta$, there is some $\varepsilon > 0$ such that . . . □

Example 1 (External forcing). Consider the system . . .

Corollary 1. *Assume H is C^2 and (a_∞, b_∞)-subquadratic at infinity. Let . . .*

Lemma 1. *Assume that H is C^2 on $\mathbb{R}^{2n}\backslash\{0\}$ and that $H''(x)$ is . . .*

Theorem 1 (Ghoussoub-Preiss). *Let X be a Banach Space and $\Phi : X \to \mathbb{R}$ . . .*

Definition 1. *We shall say that a C^1 function $\Phi : X \to \mathbb{R}$ satisfies . . .*

14 Fine Tuning of the Text

The following should be used to improve the readability of the text:

`\,` a thin space, e.g. between numbers or between units and numbers; a line division will not be made following this space

`--` en dash; two strokes, without a space at either end

`␣--␣` en dash; two strokes, with a space at either end

`-` hyphen; one stroke, no space at either end

`$-$` minus, in the text *only*

Input

```
21\,$^{\circ}$C etc.,
Dr h.\,c.\,Rockefellar-Smith \dots
20,000\,km and  Prof.\,Dr Mallory \dots
1950--1985 \dots
this -- written on a computer -- is now printed
$-30$\,K \dots
```

Output

21 °C etc., Dr h. c. Rockefellar-Smith . . .
20,000 km and Prof. Dr Mallory . . .
1950–1985 . . .
this – written on a computer – is now printed
-30 K . . .

15 Special Typefaces

Normal type (roman text) need not be coded. *Italic* (`{\em <text>}` better still `\emph{<text>}`) or, if necessary, **boldface** should be used for emphasis.

`{\itshape Text}` *Italicized Text*

`{\em Text}` *Emphasized Text – if you would like to emphasize a* definition *within an italicized text (e.g. of a* theorem) *you should code the expression to be emphasized by* `\em`.

`{\bfseries Text}` **Important Text**

`\vec{Symbol}` Vectors may only appear in math mode. The default LaTeX vector symbol has been adapted[3] to LLNCS conventions.
`$\vec{A \times B\cdot C}` yields $\boldsymbol{A} \times \boldsymbol{B} \cdot \boldsymbol{C}$
`$\vec{A}^{T} \otimes \vec{B} \otimes \vec{\hat{D}}$`yields $\boldsymbol{A}^T \otimes \boldsymbol{B} \otimes \hat{\boldsymbol{D}}$

[3] If you absolutely must revive the original LaTeX design of the vector symbol (as an arrow accent), please specify the option `[orivec]` in the `documentclass` line.

16 Footnotes

Footnotes within the text should be coded:

```
\footnote{Text}
```

Sample Input

Text with a footnote`\footnote{The footnote is automatically numbered.}` and text continues . . .

Sample Output

Text with a footnote[4] and text continues . . .

17 Lists

Please code lists as described below:

Sample Input

```
\begin{enumerate}
  \item First item
  \item Second item
  \begin{enumerate}
    \item First nested item
    \item Second nested item
  \end{enumerate}
  \item Third item
\end{enumerate}
```

Sample Output

1. First item
2. Second item
   (a) First nested item
   (b) Second nested item
3. Third item

18 Figures

Figure environments should be inserted after (not in) the paragraph in which the figure is first mentioned. They will be numbered automatically.

Preferably the images should be enclosed as PostScript files – best as EPS data using the epsfig package.

If you cannot include them into your output this way and use other techniques for a separate production, the figures (line drawings and those containing

[4] The footnote is automatically numbered.

halftone inserts as well as halftone figures) *should not be pasted into your laser-printer output.* They should be enclosed separately in camera-ready form (original artwork, glossy prints, photographs and/or slides). The lettering should be suitable for reproduction, and after a probably necessary reduction the height of capital letters should be at least 1.8 mm and not more than 2.5 mm. Check that lines and other details are uniformly black and that the lettering on figures is clearly legible.

To leave the desired amount of space for the height of your figures, please use the coding described below. As can be seen in the output, we will automatically provide 1 cm space above and below the figure, so that you should only leave the space equivalent to the size of the figure itself. Please note that "x" in the following coding stands for the actual height of the figure:

```
\begin{figure}
\vspace{x cm}
\caption[ ]{...text of caption...}        (Do type [ ])
\end{figure}
```

Sample Input

```
\begin{figure}
\vspace{2.5cm}
\caption{This is the caption of the figure displaying a white
eagle and a white horse on a snow field}
\end{figure}
```

Sample Output

Fig. 1. This is the caption of the figure displaying a white eagle and a white horse on a snow field

19 Tables

Table captions should be treated in the same way as figure legends, except that the table captions appear *above* the tables. The tables will be numbered automatically.

19.1 Tables Coded with LaTeX

Please use the following coding:

Sample Input

```
\begin{table}
\caption{Critical $N$ values}
\begin{tabular}{llllll}
\hline\noalign{\smallskip}
${\mathrm M}_\odot$ & $\beta_{0}$ & $T_{\mathrm c6}$ & $\gamma$
  & $N_{\mathrm{crit}}^{\mathrm L}$
  & $N_{\mathrm{crit}}^{\mathrm{Te}}$\\
\noalign{\smallskip}
\hline
\noalign{\smallskip}
 30 & 0.82 & 38.4 & 35.7 & 154 & 320 \\
 60 & 0.67 & 42.1 & 34.7 & 138 & 340 \\
120 & 0.52 & 45.1 & 34.0 & 124 & 370 \\
\hline
\end{tabular}
\end{table}
```

Sample Output

Table 1. Critical N values

$\mathrm{M}_\odot$	β_0	T_{c6}	γ	$N_{\mathrm{crit}}^{\mathrm{L}}$	$N_{\mathrm{crit}}^{\mathrm{Te}}$
30	0.82	38.4	35.7	154	320
60	0.67	42.1	34.7	138	340
120	0.52	45.1	34.0	124	370

Before continuing your text you need an empty line. . . .

For further information you will find a complete description of the tabular environment on p. 62 ff. and p. 204 of the *LaTeX User's Guide & Reference Manual* by Leslie Lamport.

19.2 Tables Not Coded with LaTeX

If you do not wish to code your table using LaTeX but prefer to have it reproduced separately, proceed as for figures and use the following coding:

Sample Input

```
\begin{table}
\caption{text of your caption}
\vspace{x cm}     % the actual height needed for your table
\end{table}
```

19.3 Signs and Characters

Special Signs. You may need to use special signs. The available ones are listed in the LaTeX *User's Guide & Reference Manual* by Leslie Lamport, pp. 41 ff. We have created further symbols for math mode (enclosed in $):

`\grole`	yields	$\gtreqless$	`\getsto`	yields	$\leftrightarrows$
`\lid`	yields	$\leqq$	`\gid`	yields	$\geqq$

Gothic (Fraktur). If gothic letters are *necessary*, please use those of the relevant AMS-TeX alphabet which are available using the amstex package of the American Mathematical Society.

In LaTeX only the following gothic letters are available: `$\Re$` yields $\Re$ and `$\Im$` yields $\Im$. These should *not* be used when you need gothic letters for your contribution. Use AMS-TeX gothic as explained above. For the real and the imaginary parts of a complex number within math mode you should use instead: `$\mathrm{Re}$` (which yields Re) or `$\mathrm{Im}$` (which yields Im).

Script. For script capitals use the coding

`$\mathcal{AB}$` which yields $\mathcal{AB}$

(see p. 42 of the LaTeX book).

Special Roman. If you need other symbols than those below, you could use the blackboard bold characters of AMS-TeX, but there might arise capacity problems in loading additional AMS-TeX fonts. Therefore we created the blackboard bold characters listed below. Some of them are not esthetically satisfactory. This need not deter you from using them: in the final printed form they will be replaced by the well-designed MT (monotype) characters of the phototypesetting machine.

`\bbbc`	(complex numbers)	yields	$\mathbb{C}$	`\bbbf`	(blackboard bold F)	yields	$\mathbb{F}$
`\bbbh`	(blackboard bold H)	yields	$\mathbb{H}$	`\bbbk`	(blackboard bold K)	yields	$\mathbb{K}$
`\bbbm`	(blackboard bold M)	yields	$\mathbb{M}$	`\bbbn`	(natural numbers N)	yields	$\mathbb{N}$
`\bbbp`	(blackboard bold P)	yields	$\mathbb{P}$	`\bbbq`	(rational numbers)	yields	$\mathbb{Q}$
`\bbbr`	(real numbers)	yields	$\mathbb{R}$	`\bbbs`	(blackboard bold S)	yields	$\mathbb{S}$
`\bbbt`	(blackboard bold T)	yields	$\mathbb{T}$	`\bbbz`	(whole numbers)	yields	$\mathbb{Z}$
`\bbbone`	(symbol one)	yields	$\mathbb{1}$				

$$\mathbb{C}^{\mathbb{C}^{\mathbb{C}}} \otimes \mathbb{F}_{\mathbb{F}_{\mathbb{F}}} \otimes \mathbb{H}_{\mathbb{H}_{\mathbb{H}}} \otimes \mathbb{K}_{\mathbb{K}_{\mathbb{K}}} \otimes \mathbb{M}^{\mathbb{M}^{\mathbb{M}}} \otimes \mathbb{N}_{\mathbb{N}_{\mathbb{N}}} \otimes \mathbb{P}^{\mathbb{P}^{\mathbb{P}}}$$

$$\otimes \mathbb{Q}_{\mathbb{Q}_{\mathbb{Q}}} \otimes \mathbb{R}^{\mathbb{R}^{\mathbb{R}}} \otimes \mathbb{S}^{\mathbb{S}_{\mathbb{S}}} \otimes \mathbb{T}^{\mathbb{T}^{\mathbb{T}}} \otimes \mathbb{Z} \otimes \mathbb{1}^{\mathbb{1}_{1}}$$

20 References

There are three reference systems available; only one, of course, should be used for your contribution. With each system (by number only, by letter-number or by author-year) a reference list containing all citations in the text, should be included at the end of your contribution placing the LaTeX environment `thebibliography` there. For an overall information on that environment see the *LaTeX User's Guide & Reference Manual* by Leslie Lamport, p. 71.

There is a special BibTeX style for LLNCS that works along with the class: `splncs.bst` – call for it with a line `\bibliographystyle{splncs}`. If you plan to use another BibTeX style you are customed to, please specify the option `[oribibl]` in the `documentclass` line, like:

```
\documentclass[oribibl]{llncs}
```

This will retain the original LaTeX code for the bibliographic environment and the `\cite` mechanism that many BibTeX applications rely on.

20.1 References by Letter-Number or by Number Only

References are cited in the text – using the `\cite` command of LaTeX – by number or by letter-number in square brackets, e.g. [1] or [E1, S2], [P1], according to your use of the `\bibitem` command in the `thebibliography` environment. The coding is as follows: if you choose your own label for the sources by giving an optional argument to the `\bibitem` command the citations in the text are marked with the label you supplied. Otherwise a simple numbering is done, which is preferred.

```
The results in this section are a refined version
of \cite{clar:eke}; the minimality result of Proposition~14
was the first of its kind.
```

The above input produces the citation: "... refined version of [CE1]; the minimality...". Then the `\bibitem` entry of the `thebibliography` environment should read:

```
\begin{thebibliography}{[MT1]}
.
.
\bibitem[CE1]{clar:eke}
Clarke, F., Ekeland, I.:
Nonlinear oscillations and boundary-value problems for
Hamiltonian systems.
Arch. Rat. Mech. Anal. 78, 315--333 (1982)
.
.
\end{thebibliography}
```

The complete bibliography looks like this:

References

[1] Sebastiani, Fabrizio. ”Machine learning in automated text categorization.” ACM computing surveys (CSUR) 34.1 (2002): 1-47.

[2] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, ”A Bayesian Approach to Filtering Junk E-Mail”, In Proc. of the AAAI Workshop on Learning for Text Categorization, 1998.

[3] Kiritchenko, Svetlana, Stan Matwin, and Suhayya Abu-Hakima. ”Email classification with temporal features.” Intelligent Information Processing and Web Mining. Springer Berlin Heidelberg, 2004. 523-533.

[4] Martin, Steve, et al. ”Analyzing Behavioral Features for Email Classification.” CEAS. 2005.

[5] Provost, Jefferson. ”Naive-Bayes vs. Rule-Learning in Classification of Email.” University of Texas at Austin (1999).

[6] Youn, Seongwook, and Dennis McLeod. ”A comparative study for email classification.” Advances and Innovations in Systems, Computing Sciences and Software Engineering. Springer Netherlands, 2007. 387-391.

[7] Nigam, Kamal, et al. ”Text classification from labeled and unlabeled documents using EM.” Machine learning 39.2-3 (2000): 103-134.

[8] Email classification with co-training. Svetlana Kiritchenko, Stan Matwin. In Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative Research, CASCON 2001.

[9] Analyzing the Effectiveness and Applicability of Co-training. Kamal Nigam, Rayid Ghani. In Proceedings of the Ninth International Conference on Information and Knowledge Management, CIKM 2000.