ELSEVIER

Cognitive Science 27 (2003) 159–193

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Sequential sampling models of human text classification

Michael D. Lee*, Elissa Y. Corlett

Department of Psychology, University of Adelaide, Adelaide 5005, SA, Australia

Received 21 February 2002; received in revised form 28 August 2002; accepted 17 October 2002

Abstract

Text classification involves deciding whether or not a document is about a given topic. It is an important problem in machine learning, because automated text classifiers have enormous potential for application in information retrieval systems. It is also an interesting problem for cognitive science, because it involves real world human decision making with complicated stimuli. This paper develops two models of human text document classification based on random walk and accumulator sequential sampling processes. The models are evaluated using data from an experiment where participants classify text documents presented one word at a time under task instructions that emphasize either speed or accuracy, and rate their confidence in their decisions. Fitting the random walk and accumulator models to these data shows that the accumulator provides a better account of the decisions made, and a "balance of evidence" measure provides the best account of confidence. Both models are also evaluated in the applied information retrieval context, by comparing their performance to established machine learning techniques on the standard Reuters-21578 corpus. It is found that they are almost as accurate as the benchmarks, and make decisions much more quickly because they only need to examine a small proportion of the words in the document. In addition, the ability of the accumulator model to produce useful confidence measures is shown to have application in prioritizing the results of classification decisions.

Keywords: Text classification; Sequential sampling processes; Random walks; Accumulators; Information retrieval

* Corresponding author. Tel.: +61-8-8303-6096; fax: +61-8-8303-3770.
E-mail address: michael.lee@adelaide.edu.au (M.D. Lee).

0364-0213/02/$ – see front matter © 2002 Cognitive Science Society, Inc. All rights reserved.
PII: S0364-0213(02)00117-9

1. Introduction

A central problem in information retrieval is the classification of text documents. Given a particular document, and a particular topic, the classification problem is to determine whether or not the document is about the topic. A range of machine learning techniques have been applied to the text classification problem (see, for example, Yang & Liu, 1999), many of which involve solving difficult optimization problems, or doing other extensive calculations. In addition, most of these text classifiers consider every word in the document, even when the individual documents have a large number of words. Taken together, these properties mean that machine learning classifiers take time to process large text corpora. While this is not always a problem, there are some applied situations where users require fast "on-line" text document classification for large numbers of documents.

As with many artificial intelligence and machine learning problems, there is much to be learned from examining the way in which humans perform the task of text classification. In particular, it is worth making the effort to understand how people manage to make quick and accurate decisions regarding which of the many text documents they encounter everyday—newspaper articles, e-mails, journal articles, postal correspondence, and so on—are about topics of interest. Conversely, cognitive models of human decision making can benefit from studying human performance on a real world task such as text classification, using complicated natural stimuli such as text documents. There are, of course, advantages in studying decision making with artificial stimuli, as if often done in the categorization and decision making literature (e.g., Bourne, 1974; Medin & Schaffer, 1978; Nosofsky, 1986; Shepard, Hovland, & Jenkins, 1961), because of the experimental control that is achievable. A central and long-standing argument of ecological approaches (e.g., Brunswik, 1943; Simon, 1956), however, is that it is also important to consider the role of non-arbitrary stimulus environments in supporting (or confounding) human decision making. Indeed, a number of more recent research efforts have explicitly incorporated models of environmental structure in developing formal accounts of human cognitive processes. These include Shepard's (1987, 1994) (see also Myung & Shepard, 1996) theory of stimulus generalization, Anderson's (1990, 1991, 1992) rational theory of memory, categorization, inference and problem solving, and the "fast and frugal" heuristic models of decision making developed by Gigerenzer and Todd (1999) (see also Todd & Gigerenzer, 2000).

This paper develops and evaluates cognitive models of human decision making, and also uses these models to design and test automated text classification systems. The structure of this paper is as follows: in the next section, three psychological observations about human text classification are outlined. These observations relate to the way in which people seem to make text classification decisions, and provide an impetus for developing models that use random walk and accumulator sequential sampling processes. The results of an experiment are then presented, in which people classify text documents presented one word at a time, under instruction conditions that emphasize either speed or accuracy. The ability of the random walk and accumulator models to capture the decisions made by people, their confidence in those decisions, and the number of words they read before making the decisions, is then examined. Finally, the random walk and accumulator models are evaluated as automated text classification systems, by comparing their performance to established machine learning techniques on a benchmark problem.

2. Psychological observations about text classification

2.1. Non-compensatory decision making

When people decide whether or not a text document is about a topic, they often make non-compensatory decisions, in the sense that they do not consider all of the words in the document. For example, if asked whether a newspaper article is about the U.S. Presidency, and the first seven words read are "Motorists living in rural and regional Australia . . . ," many people might choose to answer "no," even if they were permitted to read the remainder of the article.

In developing their fast and frugal heuristics, Gigerenzer and Todd (1999) present a compelling case for the role of environmental structure in facilitating non-compensatory decision making. Their basic argument is that people are able to use efficient and robust decision making strategies by relying on regularities in their task environment. If, for example, an environment has a structure where the first pieces of information found in a search are predictive of the information that would be found by more extensive searching, it is possible (and sensible) to make a decision based on a limited search. It is also reasonable to make limited searches in environments with diminishing returns, where the first pieces of information are significantly more important than those that follow. When these sorts of regularities exist in a task environment, non-compensatory decision making provides a mechanism for making decisions that are both fast and accurate.

In the context of text document classification, it seems likely that words near the beginning of a text document will often provide some clear indication of the semantic topic of that document. Generally, writers inform their audience of the topic of a document at or near the beginning of the document, and so early words should provide a strong indication of the topic of a document. If this is true, it provides an environmental regularity to which people are almost certainly sensitive, and could enable the effective use of non-compensatory decision making.

2.2. Competing models in decision making

A second psychological observation involves the relationship between the "yes" decision "the document is about the topic," and the "no" decision "the document is not about the topic." When people are asked to make this decision, they actively seek information that would help them make either choice. In other words, it is possible for both "yes" and "no" decisions to be made in a non-compensatory way.

For example, if asked whether a newspaper article is about the U.S. Presidency, and the first word is "The," it seems likely that most people would not be able to make a decision with any degree of confidence. If, however, the first word is "Clinton," it seems likely that most people would confidently respond "yes." Conversely, if the first word is "Cricket," it seems likely most people would confidently respond "no."

When people answer "no" in the final scenario, it suggests that they are actively evaluating the word "Cricket" as evidence in favor of the document not being about the topic (in the same way they actively evaluate the word "Clinton" as evidence that the document is about the topic).

This behavior implies that people treat the "yes" and "no" choices as two competing models, and are able to use the content of the document as evidence in favor of either model. Importantly, this behavior is not consistent with a "hypothesis-testing" approach to decision making, where the decision "yes" is treated as an alternative hypothesis, and is accepted if sufficient evidence is found in its favor, but is rejected in favor of the null hypothesis of deciding "no" if insufficient evidence is found.

2.3. Complete decision making

A third psychological observation is that when people decide whether or not a text document is about a topic, they generate more information than just a binary choice. People give answers having taken a period of time, and are able to express a level of confidence in their decision. As argued by Vickers and Lee (1998, p. 178), these measures provide a source of additional empirical constraints that assist in the process of model development, evaluation and comparison. Certainly, it is important that a cognitive model of the text classification decision making process is able to make predictions about performance measures such as confidence and time.

3. Sequential sampling models of text classification

Sequential sampling process models of decision making (e.g., Busemeyer & Rapoport, 1988; Laming, 1968; Link & Heath, 1975; Nosofsky & Palmeri, 1997; Ratcliff, 1978; Smith, 2000; Vickers, 1979) assume that stimuli are continually sampled for information, until sufficient evidence has been accrued to favor one decision over the alternatives, or no more information is available. Most commonly, these models involve random walk or accumulator processes. In a random walk model, where there are two alternative decisions, each successive piece of information is used to adjust an accrued evidence total, and a decision is made once a threshold level of information has been reached for one of the decisions. Accumulator models, in contrast, maintain separate evidence totals for both possible decisions, and make a decision when one of these totals reaches a threshold. Within these general frameworks, random walk and accumulator models allow all sorts of variations, involving issues such as how information is accumulated and retained, and how thresholds are regulated (see Smith, 2000 for a detailed technical discussion).

The important point is that random walk and accumulator models naturally capture the three psychological observations. Both models establish explicit evidence thresholds for each possible decision; non-compensatory decisions can be made, since the stimulus is only examined until the point where the threshold is exceeded; and both models generate predictions regarding how much stimulus information will be gathered before a decision is made, and what measure of confidence will be given to that decision.

This integration of the psychological observations suggests text classifiers that examine each word in a text document sequentially, evaluating the extent to which that word favors the alternative "yes" and "no" decisions, and using the evidence value to update the state of a random walk or accumulator model. All that is missing from this specification is a concrete formulation of the notion of evidence.

3.1. Measuring evidence

The evidence measure developed here is essentially a measure of how often a word occurs in documents about a topic relative to how often it occurs in documents not about that topic. The presence of a word like "Clinton" in a document provides strong evidence that the document is about the U.S. Presidency, because it occurs regularly in documents about the topic, and rarely in documents that are not about the topic. In contrast, a word like "Cricket" provides strong evidence against a document being about the presidency, because it seldom occurs in documents about the topic, but does appear in documents about other topics. Finally, words like "the" or "tine" provides little evidence in favor of either decision, because they occur at the same rate in documents both about and not about the topic (which is often in the case of "the," and much less often in the case of "tine").

Using these ideas, the evidence that the *i*th word in a dictionary provides about topic T, denoted by $V_T(w_i)$, may be defined formally on a log-odds scale as follows:

$$V_T(w_i) = \ln \frac{p(w_i|T)}{p(w_i|\bar{T})} \approx \ln \frac{|w_i \in T|/|T|}{|w_i \in \bar{T}|/|\bar{T}|}, \tag{1}$$

where T is "about a topic," $\bar{T}$ is "not about a topic," $|w_i \in T|$ is the number of times word w_i occurs in documents about topic T, and $|T|$ is the total number of words in documents about topic T. Because they lie on a log-odds scale, these evidence values are symmetric about zero: words with positive values ("Clinton") suggest that the document is about the topic, words with negative values ("cricket") suggest that the document is not about the topic, and words with values near zero ("the," "tine") provide little evidence for either alternative. In practice, evidence values can be calculated using a text corpus that contains a reasonably large number of documents, each of which has been identified as either being about or not being about a set of topics.

Throughout this study, the standard Reuters-21578 text corpus (Lewis, 1997) is used as a source of real world text documents. This corpus contains a set of 21,578 news articles and involves 90 topics, covering a diverse range of trade, resource and economic concepts, such as "copper," "housing," "money-supply" and "coffee." Every article has been assessed against every topic by human readers, with a list of those topics an article is judged to be about prefacing the document title, text, and other metadata in the corpus.

Beyond converting all characters to lower case, no pre-processing of the corpus, such as word stemming, was undertaken. This means that words were defined simply as unique strings of characters and numbers separated by spaces. The evidence that each such word provides for each topic was calculated according to Eq. (1), using the so-called "ModApte split" training set (Lewis, 1997; Yang & Liu, 1999). This split defines a standardized way of separating the corpus into a set of training documents and a set of test documents, and so allows independent research efforts to be compared meaningfully, because they are tackling the same basic problem.

From the evidence values, the possibility that the documents have an environmental regularity, in the form of using higher evidence words at or near their beginning, is able to be tested in a quantitative way. Fig. 1 shows the mean absolute evidence provided by words according to their relative position in the documents. It can be seen that words at the beginning of documents provide much more evidence than those in the middle or near the end, although there is a small increase for words at the very end, presumably associated with "summing up." Of course, this

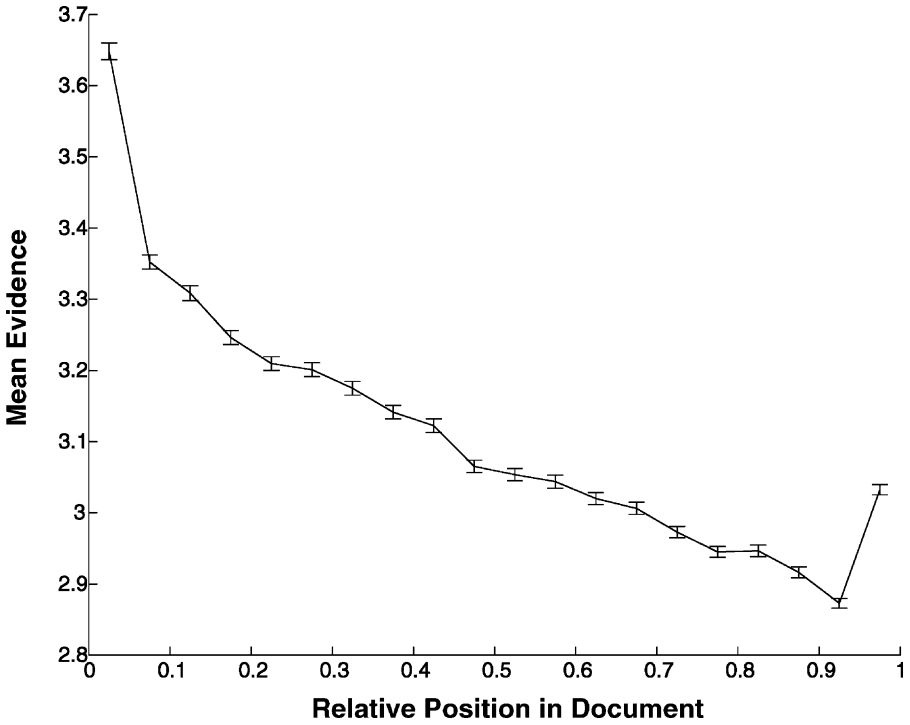

Fig. 1. The mean absolute evidence provided by words in the Reuters-21578 corpus, as a function of their relative position in the document.

analysis of one corpus does not prove that there is a general environmental regularity across all text documents, and it may well be the case that corpora from the other sorts of genre studied in the field of text and discourse processing (e.g., Biber, 1988) do not exhibit the same pattern. It would be an interesting exercise to measure the change in evidence across document position for other collections of news articles, as well as for other forms of writing, such as longer magazine-style documents, formal and informal personal correspondence, instruction manuals, James Joyce novels, speech that has been converted to text, documents written in languages other than English, and so on.

3.2. Random walk model

In random walk models, the total evidence is calculated as the difference between the evidence for the two competing alternatives, and a decision is made once it reaches an upper or lower threshold. This process can be interpreted in Bayesian terms (e.g., Carlin & Louis, 2000; Gelman, Carlin, Stern, & Rubin 1995; Leonard & Hsu, 1999; Lindley, 1972), where the state of the random walk is the log posterior odds of the document being about the topic. Using Bayes' theorem, the log posterior odds is given by

$$\ln\frac{p(T|D)}{p(\bar{T}|D)} = \ln\frac{p(T)}{p(\bar{T})} + \ln\frac{p(D|T)}{p(D|\bar{T})},$$

where D is the document being classified in terms of topic T. We then assume that the document can be represented in terms of its n words $w_1, w_2, \ldots, w_n$, so that:

$$\ln \frac{p(T|D)}{p(\bar{T}|D)} \approx \ln \frac{p(T)}{p(\bar{T})} + \ln \frac{p(w_1, w_2, \ldots, w_n|T)}{p(w_1, w_2, \ldots, w_n|\bar{T})}.$$

Finally, to make the calculation tractable, it is assumed that each word provides independent evidence, so that the log posterior odds becomes:

$$\begin{aligned} \ln \frac{p(T|D)}{p(\bar{T}|D)} &= \ln \frac{p(T)}{p(\bar{T})} + \ln \frac{p(w_1|T)}{p(w_1|\bar{T})} + \ln \frac{p(w_2|T)}{p(w_2|\bar{T})} + \cdots + \ln \frac{p(w_n|T)}{p(w_n|\bar{T})} \\ &= \ln \frac{p(T)}{p(\bar{T})} + V_T(w_1) + V_T(w_2) + \cdots + V_T(w_n). \end{aligned} \tag{2}$$

This final formulation consists of a first "bias" term, given by prior probabilities of "yes" and "no" decisions, that determines the starting point of the random walk, followed by the summation of the evidence provided by each successive word in the document.

Once the random walk has terminated, and a decision made according to whether it reached an upper or lower threshold, a measure of confidence in the decision is determined by the number of words examined. For documents that require many words to classify, confidence will be low, while for documents classified quickly using few words, confidence will be high.

Fig. 2 summarizes the operation of the random walk model on a document from the Reuters-21578 collection that is about the topic being examined. The state of the random walk is shown as the evidence provided by successive words in the document is assessed. A threshold value of 50 is shown by the dotted lines above and below. This example highlights the use of non-compensatory decision making, because the evidence accrued in reading the first 100 words of the documents led to a correct "yes" decision being made, but the final state of the random walk, when the entire document has been considered, favors a "no" decision being made.

It is clear that the random walk is very similar to the standard Naive Bayes classifier, from the field of machine learning, that has previously has been applied to the text classification problem (e.g., Nigam, McCallum, Thrun, & Mitchell, 2000; Yang & Liu, 1999). Naive Bayes classifiers are effectively random walks that always consider all of the available information, rather than requiring a target level of confidence to make a decision, and do not model the time course of decision making. The potential advantages of random walks, and other sequential sampling processes, is that their emphasis on accruing information in an ordered and temporal way means they address confidence and time performance measures, and enable non-compensatory decisions to be made.

3.3. Accumulator model

The accumulator model differs from the random walk by maintaining separate evidence totals, A_T and $A_{\bar{T}}$, for the "yes" and "no" decisions, respectively. As with the random walk model, these totals may begin at non-zero values to reflect decision bias, and then accumulate evidence by reading the words in the document. When the ith word is read, the two evidence

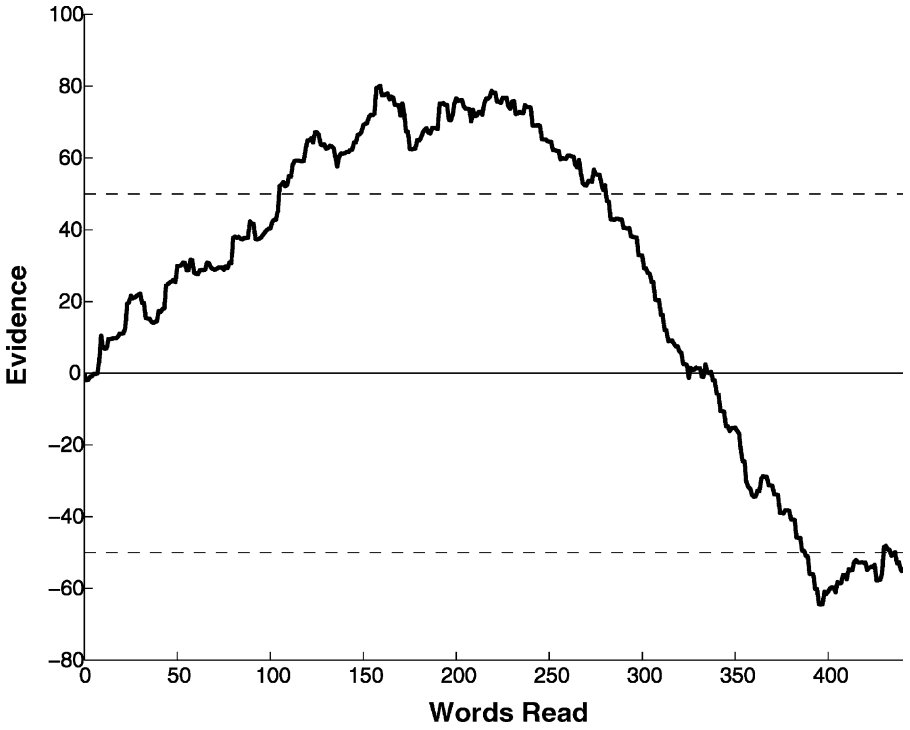

Fig. 2. Operation of the random walk model in a case where the document is about the topic.

totals are updated as follows:

$$A_T \leftarrow \begin{cases} A_T + V_T(w_i), & \text{if } V_T(w_i) > 0, \\ A_T, & \text{if } V_T(w_i) \leq 0; \end{cases}$$

and

$$A_{\bar{T}} \leftarrow \begin{cases} A_{\bar{T}} + V_T(w_i), & \text{if } V_T(w_i) < 0, \\ A_{\bar{T}}, & \text{if } V_T(w_i) \geq 0. \end{cases}$$

In effect, this means that the evidence provided by each successive word $V_T(w_i)$ is added to the "yes" accumulator if it is positive or the "no" accumulator if it is negative. Once either the "yes" accumulator reaches a positive threshold, or the "no" accumulator reaches a negative threshold, the corresponding decision is made.

Accumulators can also be interpreted in Bayesian terms,[1] by considering each accumulator as a separate Naive Bayes classifier. As each word is read, the "yes" accumulator accrues the evidence a word provides for the document being about the topic $p(w|T)$ against the competing model given by $\min(p(w|T), p(w|\bar{T}))$. Meanwhile, the "no" accumulator accrues the evidence a word provides for the document not being about the topic $p(w|\bar{T})$ against the same competing model $\min(p(w|T), p(w|\bar{T}))$.

Because accumulators maintain separate evidence totals, there are a number of ways in which the confidence in a decision may be assessed. Following the approach used for the random

walk, confidence may be determined by the number of words read. Alternatively, confidence may be assessed using a "balance of evidence" approach (Vickers, 1979; see also Van Zandt, Colonius, & Proctor, 2000), where it is measured as the difference between the evidence totals as a proportion of the total evidence accumulated. When accumulators use asymmetric thresholds (that is, when the values of the thresholds are different), it is necessary to express the evidence in each accumulator as a proportion of its threshold (Vickers, 1985, 2001b). Formally, this means that the general balance of evidence approach calculates confidence as:

$$\frac{|A_T/k_T - |A_{\bar{T}}/k_{\bar{T}}||}{A_T/k_T + |A_{\bar{T}}/k_{\bar{T}}|}, \tag{3}$$

where k_T is the threshold for the "yes" accumulator, and $k_{\bar{T}}$ is the threshold for the "no" accumulator.

Fig. 3 shows the operation of the accumulator model on the same text document considered in Fig. 2. The state of both accumulator totals are shown as successive words in the document are read, and thresholds of 50 are once again indicated by dotted lines. As with the random walk model, the accumulator model makes a non-compensatory "yes" decision, because the "yes" accumulator is the first to reach its threshold. After all of the words in the document have been read, however, a "no" decision is favored, because the "no" accumulator has greater (absolute) evidence.

Fig. 3 also demonstrates the way in which the different methods of assessing confidence for accumulator models may make different predictions. If confidence is determined by the number of the words read, only the point at which the "yes" accumulator reaches the threshold is important. Any text document classified after the same number of words would have

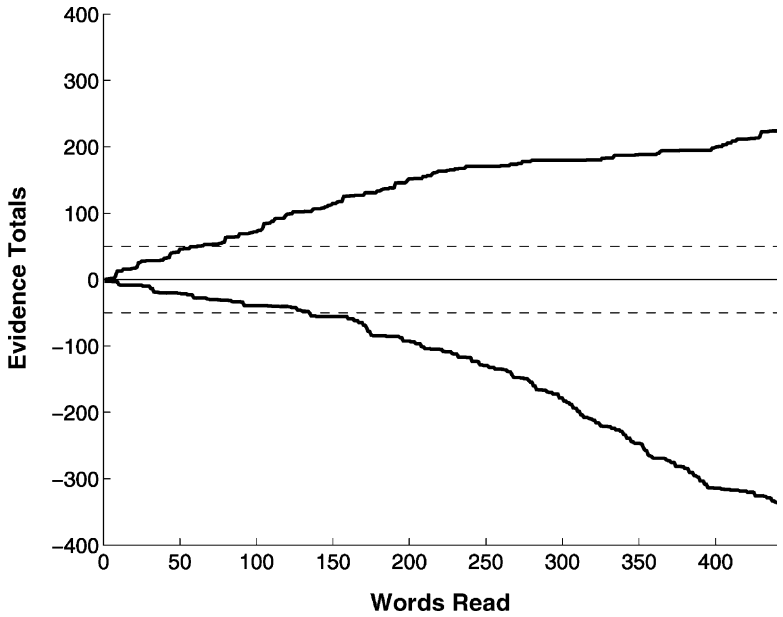

Fig. 3. Operation of the accumulator model on the same document shown in Fig. 2.

the same confidence measure. Under the balance of evidence approach, however, documents classified at the same point could be given different confidence measures depending on the level of evidence in the "no" accumulator. For example, if the "no" accumulator contains no evidence, confidence will be high, whereas, if it contains almost as much evidence as the "yes" accumulator, confidence will be low.

4. Experiment

The experiment reported here considers a limited form of human text classification, where words are presented serially at a constant rate until a decision is made regarding whether or not the document is about a given topic. While this task clearly does not incorporate important aspects of everyday text classification, such as the layout of the text, it is a task that people are able to do easily, and corresponds to the real world task of, for example, deciding whether or not the news flash on a scrolling electronic display is worth continuing to read. More importantly, the control afforded by the serial presentation methodology allows for the collection of empirical data that support quantitative evaluation of the random walk and accumulator text classification models.

4.1. Participants

Eighty-two participants, some of whom received partial course credit for their involvement, completed the text classification task. There were 51 females and 31 males, aged between 18 and 65, with a mean age of 30 years.

4.2. Stimuli

Each participant classified a total of 50 text documents. These documents, and the topics against which they were classified, were selected from the ModApte test set of the Reuters-21578 corpus. The 50 documents comprised five sets of 10 documents, where each set displayed a particular qualitative form of serial evidence accrual. The first set contained documents that, using the evidence totals learned from the training set, were observed to be consistently about the topic in question. The second set contained documents that were consistently not about the topic. The third set contained documents that started off being about the topic, but then showed a change towards not being about the topic. The fourth set contained documents that started off not being about the topic, but then showed a change towards being about the topic. These document and topic combinations proved very difficult to find, as was expected given the overall evidence structure of the corpus summarized in Fig. 1. Finally, the fifth set contained documents that never provided substantial evidence that they were either about or not about the topic.

These five qualitative forms are graphically characterized in Fig. 4, which shows the random walk pattern of evidence accrual for one of the documents against its topic for each type. For each curve, the number of words read progresses along the x-axis, and the total evidence for the alternative decisions is shown on the y-axis. The 50 documents and their topics, which are

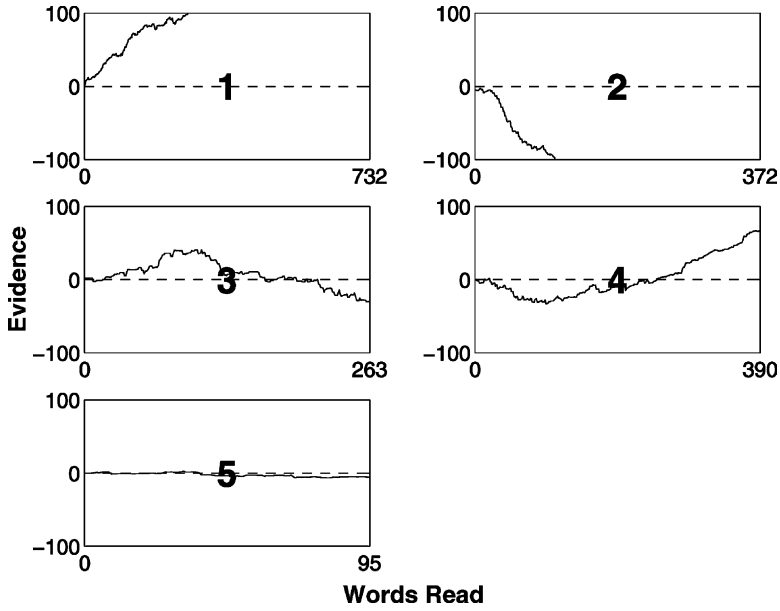

Fig. 4. Examples of each of the five question types, displayed using the random walk approach to evidence accrual.

detailed in Table 1, were selected by the visual examination of a large number of random walks like these. It is important to note that the pattern of evidence accrual was the sole criterion for selection. In particular, the assessment of human readers provided as part of the Reuters-21578 corpus was not used in any way as a basis for choosing stimuli.

Because each of the stimuli are real text documents, it is not possible to exercise precise control over the lengths of documents in each of the five sets. Table 2 provides descriptive statistics summarizing the length properties of the documents chosen, and indicates that the documents in sets 1, 2, 3, and 4, all have broadly similar length characteristics. The documents in the fifth set, however, are clearly shorter, because it proved impossible to find lengthy documents that never accrued substantial evidence for either classificatory decision.

4.3. Method

The 82 participants were randomly allocated to two experimental groups with different task instructions. In the "speed" experimental group, participants were instructed to make their decisions as quickly as possible, while participants in the "accuracy" experimental group were told to ensure their decisions were as accurate as possible.

At the commencement of each trial, the participant was presented with a question of the form "Is this document about *xxx*?," where *xxx* denoted the short description of the topic given in Table 1. The participant then pressed a "start" button, following which the body of the document (not including the document title) was presented one word at a time at a rate of one word per second, with the display of each successive word replacing the previous word. If the end of the document was reached, the text "⟨end-of-document⟩" was displayed indefinitely.

Table 1
The 50 documents and their topics used in the experiment, giving the question type, the number of the document within the Reuters corpus, the title of the document, and the topic description

Type	Number	Title	Topic
1	14826	Asian exporters fear damage from U.S.–Japan rift	Trade
1	14833	Indonesia sees CPO price rising sharply	Palm oil
1	14839	Australian foreign ship ban ends but NSW ports hit	Commercial shipping
1	20787	U.S. may end additional sanctions against Japan	Trade
1	17479	Coniston group to continue bid for Allegis	Company acquisition
1	17492	Gas carrier escaped gulf attack last week—Lloyds	Commercial shipping
1	18835	European community criticizes U.S. trade measures	Trade
1	19000	Joint action said vital to boost world growth	Gross national product
1	19028	Fed data suggest no change in monetary policy	Interest rates
1	18752	Steel firms study USX unit price hike	Iron/steel
2	18776	Indonesian debt service ratio peaks, minister says	Unemployment
2	14827	Pilots propose wage cut to fund buyout	Personal income
2	14830	Nissan may supply parts to Mexican Ford, Chrysler	Palladium
2	20778	Artillery shells said to fall on Kuwait border	Malaysian ringgit
2	14831	New Zealand raises foreign investment threshold	Wool
2	14942	Asian shows mixed performance in 1986	Linseed oil
2	18781	Growth of palm oil use set to slow, output to rise	Cotton oil
2	14824	Yugoslav workers may be angered by lost subsidies	Personal income
2	15471	Belgium launches bonds with gold warrants	Wool
2	17380	West German tapioca use seen declining	Tapioca
3	15549	U.S. said to view G-7 meeting as major success	Japanese yen
3	14923	French traders forecast EC sugar tender	Barley
3	15511	Physi-technology sees loss, in default	Company earnings
3	15567	India foodgrain target 160 million tonnes in 1987/1988	Rice
3	15219	Talks continue on tin agreement extension	Tin
3	15817	Consensus seen on tin pact extension	Tin
3	15213	Medtronic sees 15 pct earnings growth	Company earnings
3	15743	London eurodollar bonds close lower	Foreign exchange
3	14908	South Africa mines body sees May day work stoppage	Gold
3	14840	Indonesian commodity exchange may expand	Vegetable oil
4	15352	Deficit cuts seen unable to cure trade deficit	U.S. dollar
4	17101	Venezuela re-establishes posted product prices	Heating oil
4	18996	OECD urges action to cut U.S. budget deficit	Balance of payments
4	19047	U.S. seeks Japan help in event of 1988 recession	Interest rates
4	15240	Hartmarx targets earnings growth	Gross national product
4	15829	Royal Dutch unit to cut heating oil price	Heating oil
4	18834	First Wisconsin adds loan losses	Company earnings
4	16636	Texacosays some oil flows re-established	Crude oil
4	15389	RTZ sees rising U.S. output aiding 1987 results	Company earnings
4	15738	U.S. to push strong summit agriculture statement	Grain
5	14825	French government wins confidence vote	Trade
5	15033	Zambia does not plan retail maize price hike	Grain
5	16012	Egypt seeking 500,000 t corn—U.S. traders	Rice
5	20785	Pacific stock exchange closing figures delayed	Trade
5	18496	Austrian current surplus grows in 4 months	Interest rates
5	14922	Rain boosts central Queensland sugar cane crop	Coffee
5	18001	Brazil's Sarney renews call for war on inflation	Crude oil
5	17384	EC unemployment falls below 17 million in March	Wheat
5	17049	Qantas to buy extended range Boeing 767 aircraft	Company acquisition
5	17825	Fluorocarbon completes acquisition	Petroleum chemicals

Table 2
Descriptive statistical summary of document lengths across the entire stimulus set, and in terms of the five document types

	Mean	Standard deviation	Range
All	269	210	20–831
Type 1	307	219	42–732
Type 2	239	136	84–457
Type 3	328	224	98–809
Type 4	381	256	51–831
Type 5	89	50	20–181

At any time during this presentation process, the participant could make a classificatory decision by pressing a "yes" or a "no" button located immediately beneath the text. At this point, the participant was required to express their confidence in their decision. This was done using a five-point rating scale, ranging from 0, which was labeled "uncertain," through to 4, which was labeled "definitely yes" for "yes" decisions and "definitely no" for "no" decisions. Participants were instructed that, if they made a mistake in registering their decision, they should select a confidence rating of 0. The decision, confidence in the decision, and the number of words read to make the decision were then recorded, and the topic question for the next trial was presented. To control for any order effects, the 50 text documents were presented in a random order for each participant.

4.4. Results

The primary emphasis of our data analysis involves fitting the random walk and accumulator models, which is done in the next section. For this reason, we only provide a brief analysis of the raw data here, with a focus on what information the human performance data provide about the usefulness of sequential sampling accounts. We also try to achieve some clarity by reporting only a few statistical inference results drawn from a more extensive analysis, selecting those providing information that is not visually obvious from displaying the data. Our statistical inferences take the form of Bayes factors (BF: Kass & Raftery, 1995), comparing the odds that two groups of scores came from the same distribution as opposed to two separate distributions.[2]

4.4.1. Decisions

The left panel of Fig. 5 shows the relationship between the classification decisions for each document across the speed and accuracy instruction conditions. Each point represents one of the 50 documents. The x-coordinate indicates the relative number of "yes" and "no" decisions under speed instructions, while the y-coordinate indicates the relative number of "yes" and "no" decisions under accuracy instructions. One standard error around these mean decisions is shown for both experimental groups. The results suggest that, for most of the documents, the speed and accuracy task instructions did not affect the relative proportion of "yes" and "no" decisions. The right panel of Fig. 5 shows the relationship between decisions aggregated across each document type, and suggests that each type, except perhaps Type 1, gave the

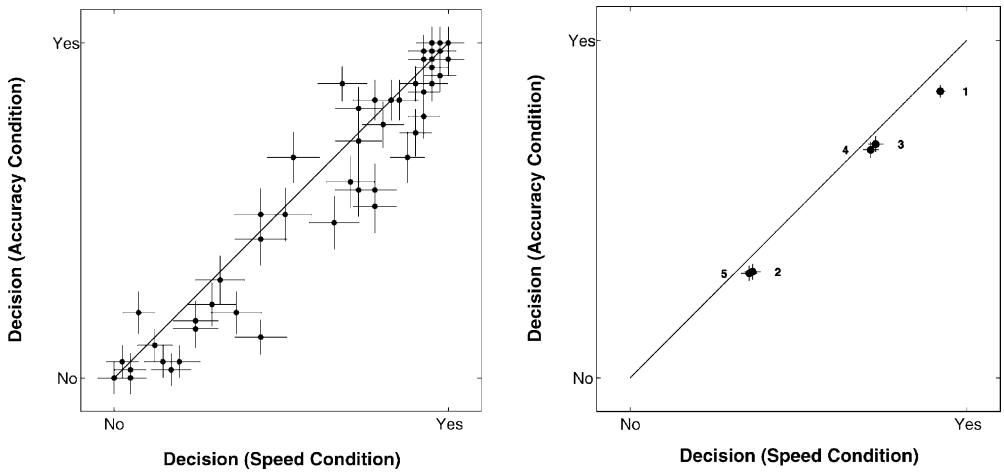

Fig. 5. The mean decision for each document (on the left), and for each document type (on the right), in terms of both speed and accuracy conditions. One standard error is shown about the means.

same decisions across task instructions. Statistical inference supports these conclusions. For Type 1 documents, it is almost 13 times more likely that the speed and accuracy condition decisions have different underlying rates of "yes" and "no" decisions (BF = 12.8) but, for all other document types it is about five times more likely that the speed and accuracy condition decisions have the same "yes" and "no" rates (BFs = 4.1, 6.5, 6.0, and 5.0, respectively).

4.4.2. Confidence

The left panel of Fig. 6 shows the relationship between the confidence in decisions for each document across the speed and accuracy instruction conditions. As with decisions, the different

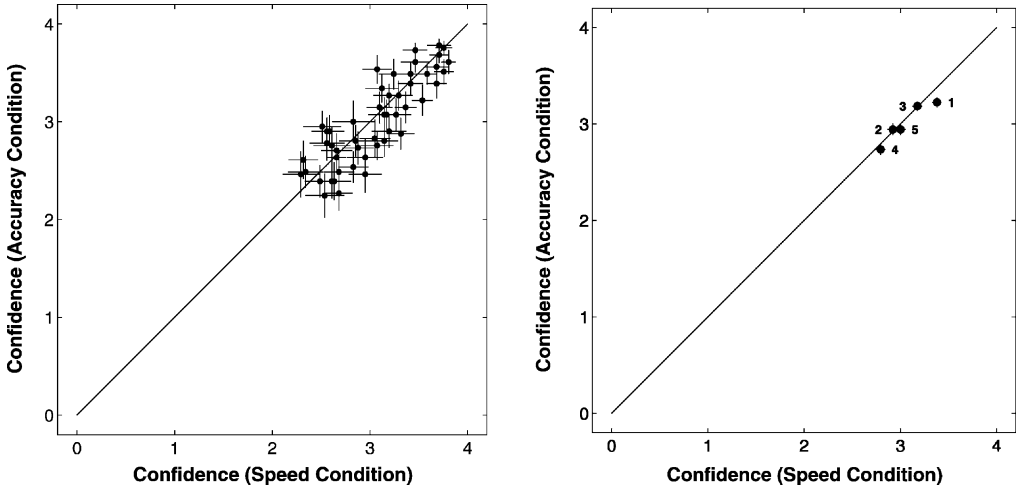

Fig. 6. The mean confidence for each document (on the left), and for each document type (on the right), in terms of both speed and accuracy conditions. One standard error is shown about the means.

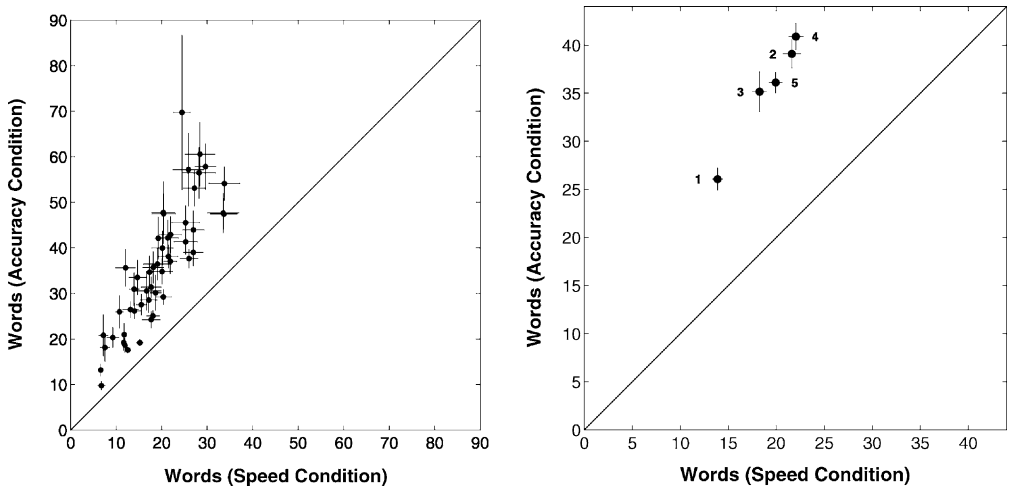

Fig. 7. The mean number of words read for each document (on the left), and for each document type (on the right), in terms of both speed and accuracy conditions. One standard error is shown about the means.

task instructions generally did not affect confidence. The right panel of Fig. 6, which shows the relationship between confidence values for each document type, suggests that confidence is the same across task instructions in all cases. Once again, statistical inference supports these conclusions, with Bayes Factors showing it is most likely the confidence scores from the speed and accuracy conditions came from Gaussian distributions with the same mean (BFs $= 2.8$, 26.5, 21.0, 16.4 and 16.4, for the five types, respectively). There are significant differences in confidence across the types, however, with Type 5 being different from Type 3 (BF $= 28.6$) and Type 4 (BF $= 17.0$), Type 2 being different from Type 3 (BF $= 152.3$) and Type 4 (BF $= 3.3$), and Type 1 being different from Type 3 (BF $= 12.9$). The only comparison where it is more likely confidence is the same is between Types 2 and 5 documents (BF $= 12.9$).

4.4.3. Number of words read

The left panel of Fig. 7 shows the relationship between the number of words read in each document across the speed and accuracy instruction conditions. The effect of the different instructions on the number of words read is clear, with participants always reading more words, on average, in a document under the accuracy condition. The right panel of Fig. 7 shows the relationship between the number of words read for each document type. The best-fitting line through the origin and the five points explains 98% of the variance and has a slope of 1.85. This suggests that the effect of the accuracy instructions is proportionately the same for each document type, and corresponds to participants reading about 85% more words before classifying a document when under accuracy instructions. It is also clear that Type 1 documents are classified using fewer words in both the speed and accuracy conditions.

4.4.4. Non-compensatory decision making

Fig. 8 examines the number of words read in terms of both instruction conditions, and the actual decision made. The top panel shows the distributions of words read for “yes” and “no”

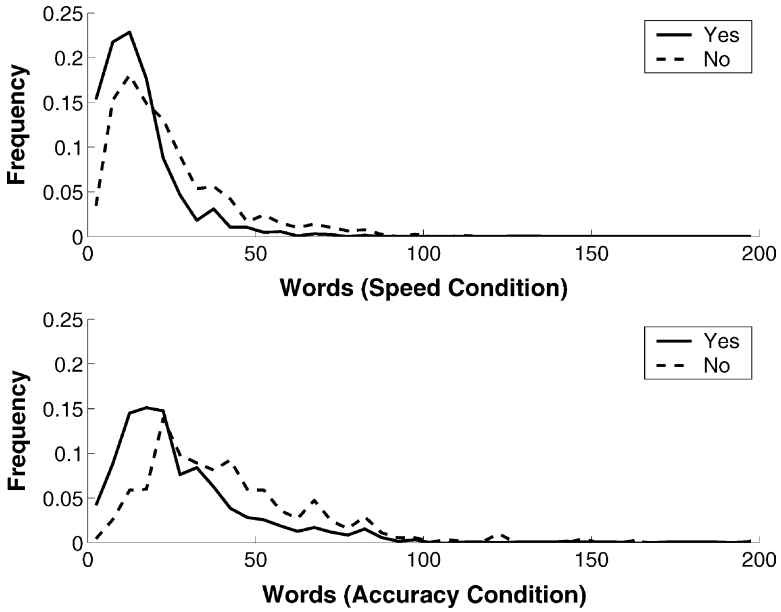

Fig. 8. The distribution of the number of words read to make yes and no decisions, for the speed condition (top panel), and the accuracy condition (bottom panel).

separately under speed instructions, and shows that "yes" decisions tend to be made using fewer words than "no" decisions. It is also clear, however, that almost all documents were classified by all participants within about the first 60 words, regardless of the decision made. The bottom panel, which relates to the accuracy condition of task instruction, shows a similar pattern: "yes" decisions are made using fewer words than "no" decisions, but almost all decisions are made within about 100 words. These distributions make it clear that the participants made both "yes" and "no" decisions in a non-compensatory way. In fact, of the $82 \times 50 = 4{,}100$ decisions made in total, there are only three cases where a participant read to the end of the document.

4.4.5. Interval of uncertainty

A theoretical device sometimes used in sequential sampling models (e.g., Juslin & Olsson, 1997; Vickers, 2001a; Vickers & Pietsch, 2001) is the "interval of uncertainty," which assumes that evidence values in a small range about zero are not accumulated at all. One way to test whether such an interval is operating for text classification is to examine the distribution of evidence values for the words immediately preceding a decision, as compared to the evidence distribution of all words read. In this comparison, an interval of uncertainty would be revealed by the terminating words having evidence values only outside an interval around zero. Fig. 9 shows the evidence distribution of the last words read by participants before a decision was made, and the evidence distribution of all words read. These results do not show any interval of uncertainty, since both terminating distributions have many evidence values around the value zero. This is an interesting finding, worthy of further investigation. In particular, it would be worthwhile examining what role, if any, the structure of sentences played in determining when decisions were made.

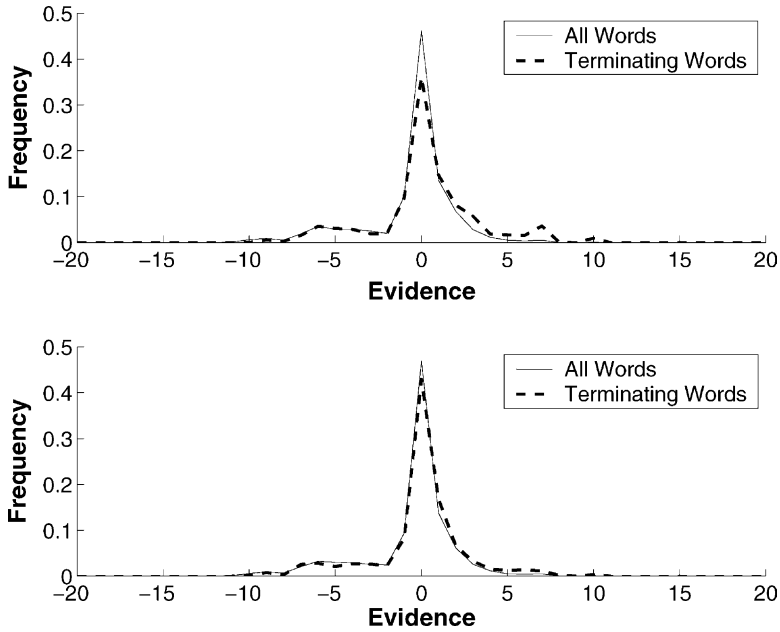

Fig. 9. The distribution of evidence values for all words read by participants (solid line), and those words immediately preceding a decision being made (broken line). Results for the speed condition are shown in the top panel, and results for the accuracy condition are shown in the bottom panel.

4.4.6. Asymmetric thresholds

The top panel of Fig. 9, which relates to the speed condition, shows that the last word read has a large positive evidence values more often that would be expected from the overall distribution.[3] No such discrepancy is evident in the lower panel of Fig. 9, which relates to the accuracy condition. This pattern of results is consistent with a sequential sampling model that combines a small "yes" threshold with a larger "no" threshold under the speed condition, and uses larger (although still possibly asymmetric) thresholds for both decisions in the accuracy condition.

4.5. Conclusions

The results of the experiment have many encouraging consistencies with a sequential sampling process account of human text classification, at least for the type of stimulus presentation used in the experiment. It is clear that people's text classification decisions are non-compensatory, and it also seems to be the case that their decision making does not suffer from a failure to read all of the words. When under accuracy conditions, participants basically made the same decisions with the same confidence as under speed conditions—they just read almost twice as many words. This suggests that decision making is not necessarily subject to a speed–accuracy tradeoff in the sense that it is improved by the consideration of more data. It seems possible to be both quick and accurate in the way advocated by Gigerenzer and Todd (1999), and implemented by random walk and accumulator models.

The pattern of decisions, words read, and confidence values across the five document types is often consistent with the definition of those types in evidence accrual terms. For example, participants made decisions more quickly when the evidence was consistently in favor of the document being about the topic, and were least confident when classifying those Type 4 documents that violated an expected environmental regularity. There are also, however, some less intuitive findings that, if replicable, would seem to need a more elaborate sequential sampling account than is considered here. In particular, it is not clear why decisions for Type 5 documents were similar to Type 2 documents, but different from Types 3 and 4. Given that Type 5 documents had patterns of evidence accrual that showed no strong evidence in favor of either decision, it might have been expected they would be classified similarly to ambiguous Types 3 and 4 documents, and differently from Type 2 documents, where the evidence is clearly in favor of "no" decisions. One possibility, to which we return in Section 7 is that, when faced with the lack of decisive information in Type 5 documents, people dynamically adjust their required evidence thresholds to enable a non-compensatory decision to be made, and so their classifications are not well described by the fixed threshold models considered here.

Meanwhile, a final important conclusion from the experimental results relates to the evidence distribution of terminating words. This distribution suggests that the inclusion of an interval of uncertainty is not necessary to model human decision making, but that people may use different standards of evidence for making "yes" and "no" decisions, and so sequential sampling models need to allow for asymmetric thresholds.

5. Fitting the empirical data

In this section, two parameter random walk and accumulator models are fit to the empirical data, where the parameters correspond to the (potentially asymmetric) evidence thresholds for "yes" and "no" decisions. Before presenting the results of this analysis, it is worth discussing three modeling challenges that are responsible for the model fitting process that is used.

5.1. Fitting process

The first modeling challenge arises because, when participants classified the documents, their decisions were far from unanimous. As Fig. 5 shows, for both conditions of task instruction, there are many documents for which different participants made different classification decisions. Presumably, these differences relate to varying interpretations of the meaning or scope of the word "about" in the question that was asked. Informal feedback suggested, for example, that some participants treated the question "Is this document about trade?" as including any document that involved goods or services being transferred between countries or regions, while others limited "yes" decisions to only those documents that met a much narrower definition, such as negotiations on bilateral trade agreements. It is beyond the scope of our models to account for these differences in semantic interpretation.

One practical means of addressing this problem is to consider only those classification decisions that were consistent across all participants. For the speed condition, there are 16 documents (11 "yes" decisions and 5 "no" decisions) where 90% or more of the participants

agreed in their decision, and there are 18 documents (10 "yes" decisions and 8 "no" decisions) that meet this criterion for the accuracy condition. In all of the modeling reported here, only these subsets of documents are used.

We should emphasize that we do not have any reason to believe that the restriction on the documents considered in model fitting favors one model over the other, and this is certainly not the intention of the manipulation. The intention is to ensure that the evidence values for each word in relation to each topic, learned using the Reuters-21578 training set, have some meaningful relationship to the internal evidence values assumed to be used by our participants. Limiting the documents considered to those where people agreed strongly in their decisions is intended to find those decisions less affected by differences in semantic interpretation, and so make it more likely that the evidence values being supplied to the random walk and accumulator processes convey the same information that drives human decision making. There is, of course, no way in which evidence values supporting a "yes" decision can (or should) be used by either model to produce a "no" decision, yet this is what would be required to explain human performance where decisions are not largely unanimous. Fundamentally, the restriction of documents to those with agreed decisions serves to give the random walk and accumulator models some (equal) chance at explaining human performance. We should acknowledge, however, that using agreed decisions does sample from the 50 questions in a biased way. Although at least one document of every type is included in the speed condition subset, there are no Type 4 documents in the accuracy condition subset. In addition, the sampling increases mean confidence from 3.1 to 3.4 for both speed and accuracy conditions, and decreases the mean number of words read from 18.0 to 14.6 for the speed condition, and from 32.9 to 25.8 for both the accuracy condition.

The second modeling challenge arises from the multiple measures of human performance provided by the decision, confidence and number of words read data. One established practice for fitting models to empirical data containing several dependent variables is to minimize the total deviation across all of the measures. This would require finding thresholds that simultaneously minimize the difference between model predictions and observed data for decisions, confidence, and number of words read. Unfortunately, the relationship between the three measures means that this standard approach does not make sense. A model, for example, that is able to predict exactly the confidence and the number of words read, but does so for the wrong decision, has fundamentally failed to provide a useful account of human performance, even though it may correspond to the minimum of an aggregated error function. In other words, models of confidence and words read presuppose an effective model of decision making. Accordingly, the approach to model fitting adopted here is to focus firstly on correctly predicting the decisions made by people, and only then seek to optimize parameter values to provide the best possible accounts of their confidence and the number of words read.

The third modeling challenge is that both random walk and accumulator models may fail to make a decision, particularly if large thresholds are used. To allow the models to be compared with the human data, where people were required to make a decision, we use forced choice versions of the models. For the random walk model, this means that, if no decision threshold has been reached at the end of the document, a classification decision is made according to the sign of the final state of the random walk. If the evidence total is positive, a "yes" decision is assumed, while a "no" decision is assumed if the total is negative. For the accumulator model, a forced choice decision is made according to the relative levels of evidence in the two evidence

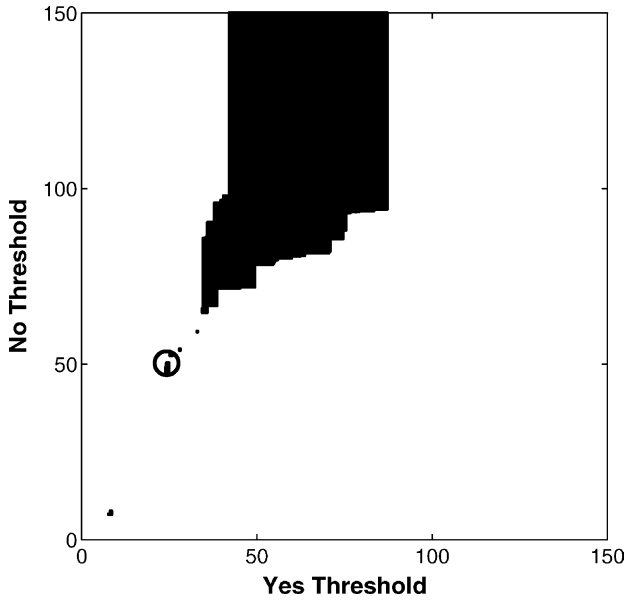

Fig. 10. The parameter space for the accumulator model, in relation to the speed data. Regions of the parameter space where the model correctly predicts all of the classifications decisions are shown in black, and the region with the best correlation with the number of words read is circled.

totals. If the "yes" evidence total exceeds the "no" evidence total a "yes" decision is made, otherwise a "no" decision is made.

5.2. *Accumulator model*

In fitting both of the models, we considered a parameter grid that extended from 0 to 500 for both "yes" and "no" thresholds, evaluated at increments of 0.25. For the accumulator model, a large number of these parameterizations matched all of the participants' decisions for both the speed and accuracy conditions. These parameterizations are shown for the speed conditions in Fig. 10, and for the accuracy condition in Fig. 11, with the correct regions of the parameter space shown in black.

Fig. 10 shows that, once the "no" threshold exceeds about 100, correct decisions are made for "yes" thresholds in the approximate range of 50–100 for the speed data. The unboundedness of this region results from the forced choice nature of the decision making, and the types of the documents in the agreed subset. There are also smaller isolated regions of the parameter space where all decisions are made correctly. In particular, using the notation ⟨"yes" threshold–"no" threshold⟩, Fig. 10 shows parameterizations at approximately 8–8, and around 25–50. Almost all of the parameterizations that lead to correct decisions use a "yes" decision threshold that is smaller than the "no" decision threshold, suggesting that participants required relatively less evidence to make a "yes" than a "no" decision.

Of the correct parameterizations, the best prediction of the number of words read is around 25–50, as shown by the circle in Fig. 10, where the correlation with human performance

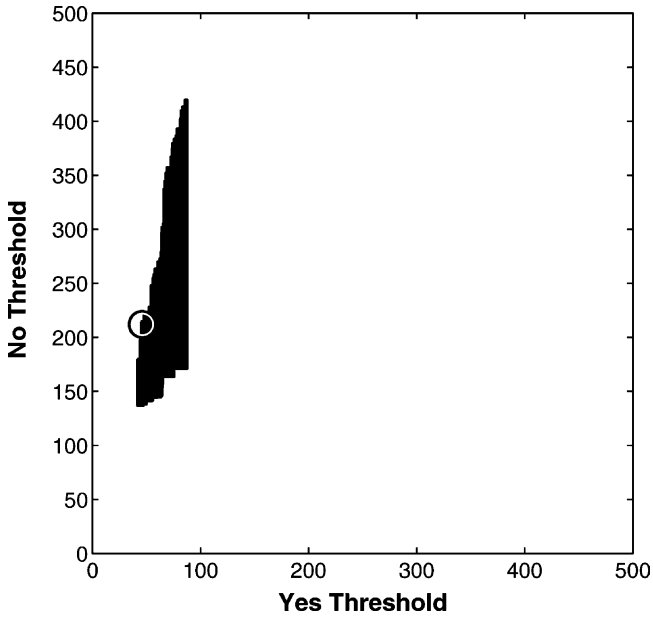

Fig. 11. The parameter space for the accumulator model, in relation to the accuracy data. Regions of the parameter space where the model correctly predicts all of the classifications decisions are shown in black, and the region with the best correlation with the number of words read is circled.

across all documents is $r = .51$. Around 25–50, the balance of evidence confidence measure correlates about $r = .65$ with human performance, whereas the words read confidence measure correlates only about $r = .32$. These confidence correlations are close to the best achieved at any of the parameterizations shown in Fig. 10: the balance of evidence confidence has a maximal correlation of $r = .67$, while the words read measure of confidence achieves a maximum of $r = .41$. Considering parameter values that achieve good correlations for both the confidence and words read suggests that parameterizations of about 25–50 provide the best account of human performance under the speed condition.

Fig. 11 shows the bounded region of the parameter space where the accumulator model makes correct decisions for the accuracy data. As with the speed data, the parameterizations all use "yes" decision thresholds that are smaller than the "no" decision thresholds. Unlike the speed case, however, there are no suitable parameterizations that use small threshold values, with the combination of approximately 50–140 being the first at which all of the documents are classified correctly. This suggests that participants under accuracy conditions were more conservative, in the sense that they required relatively greater (but still asymmetric) evidence totals for both "yes" and "no" decisions.

Of the parameterizations shown in Fig. 11, the best correlation with the number of words read is $r = .65$ at a small region around 46–212, which is circled. Parameterizations in this region have correlations with confidence of about $r = .34$ for the balance of evidence measure, and about $r = .17$ for the words read measure. Once again, these confidence correlations are not very different from the best achieved across all of the parameterizations: for the balance of

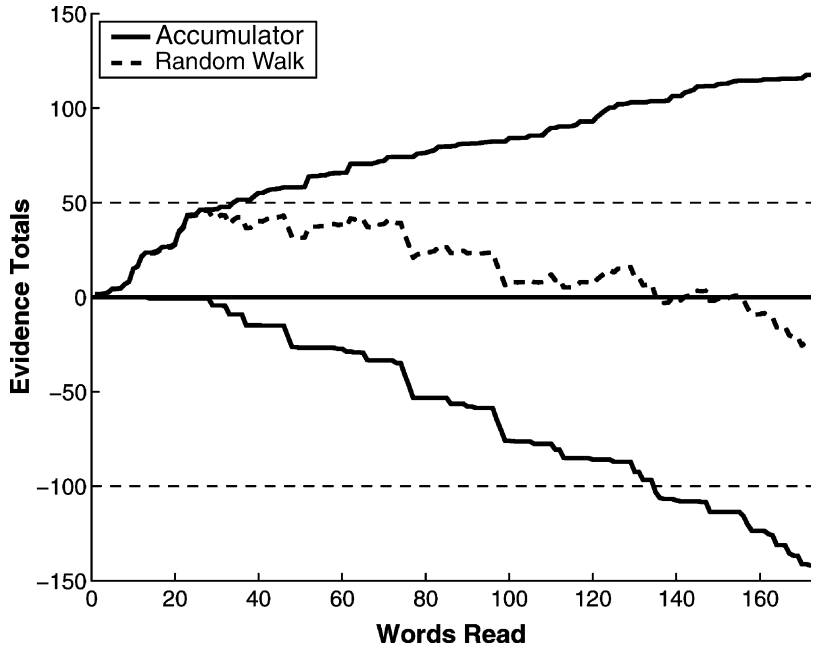

Fig. 12. The behavior of the random walk and accumulator classification models on a Type 3 document, when using asymmetric thresholds.

evidence measure, the best confidence correlation is $r = .43$, while for the words read measure, the best confidence correlation is $r = .17$.

The large number of correct parameterizations shown in Figs. 10 and 11 indicate that the accumulator model has a robust ability to capture the human decisions. This is important from a model selection perspective (e.g., Myung & Pitt, 1997; Pitt, Myung, & Zhang, 2002; Roberts & Pashler, 2000), because it shows that the fit between the model and data does not rely on a precise tuning of parameter values, and so may more confidently be attributed to the model itself.

5.3. Random walk model

At none of the parameterizations examined did the random walk model correctly predict the participants' decisions for either the speed or accuracy data. An analysis of those documents where errors were most common identified a basic deficiency in the random walk model. This deficiency is highlighted in Fig. 12, which shows the behavior of the random walk and accumulator models on a Type 3 document. For concreteness, asymmetric thresholds of 50 and 100 are shown. The important difference between the two models is that the accumulator makes a "yes" decision, because the "yes" accumulator is the first to reach the threshold, whereas the random walk model never reaches the positive threshold, and eventually makes a forced "no" decision when all of the words in the document have been read. Human classifications of these sorts of documents, as suggested by Fig. 5, strongly favor the "yes" decision, and so are consistent only with the prediction of the accumulator model.

The deficiency arises because random walks treat evidence in favor of alternative decisions as being more commensurable than accumulators, in the sense that the presence of evidence in favor of one decision can be directly negated by equal strength evidence in favor of its alternative. For example, if a document being assessed in relation to the topic "U.S. Presidency" contains a high evidence "yes" decision word like "Clinton," followed by a high evidence "no" decision word like "cricket," the random walk has little net evidence in favor of either decision, whereas the accumulator has significant (and approximately equal) evidence in favor of both. This means that, when another high evidence "yes" word appears in the document (or the word "Clinton" is repeated), an accumulator may make the "yes" decision where the random walk will not.

This is the difference in behavior highlighted in Fig. 12. The first 25 words or so all provide evidence that the document is about the topic, but the remaining words provide more ambiguous evidence, some favoring a "yes" decision and some favoring a "no" decision, with the overall effect of suggesting the document is not about the topic. For the random walk model, this net effect causes a "no" decision once all the words have been read. For the accumulator model, however, those remaining words that do suggest the document is about the topic are sufficient to prompt a "yes" decision before the "no" accumulator has gathered enough evidence. The important point is that people also make "yes" decisions when presented with these sorts of documents, presumably because seeing a word like "cricket" does not eliminate the effect of the earlier word like "Clinton." As a consequence, the evidence accrual process used by the accumulator seems better suited than the random walk to modeling the decisions made by humans in classifying text documents. It is possible, of course, that more sophisticated sequential sampling models based on random walks (e.g., Link & Heath, 1975; Ratcliff, 1978) could fit the decision data. What the analysis of presented here suggests, through its use of the most basic random walk and accumulator models, is that any such modeling success would not be attributable directly to the assumption of random walk evidence accrual.

5.4. Individual differences

Because each participant made decisions about 50 documents, there are enough data to examine individual differences. This was done by fitting separate "yes" and "no" thresholds for each participant using the accumulator model. For the speed participants, the best individual correlations with the number of words read ranged from $r = -.09$ to $r = .73$ with a mean of $r = .41$ ($SD = 0.18$). The correlations with confidence at the parameter values where the best correlations with the number of words read were achieved ranged from $r = -.05$ to $r = .79$ with a mean of $r = .59$ ($SD = 0.17$) for the balance of evidence measure, and from $r = -.63$ to $r = .39$ with a mean of $r = .09$ ($SD = 0.23$) for the words read measure. For the accuracy participants, the best individual correlations with the number of words read ranged from $r = .06$ to $r = .74$ with a mean of $r = .31$ ($SD = 0.14$). The correlations with confidence at the parameter values where the best correlations with the number of words read were achieved ranged from $r = .27$ to $r = .84$ with a mean of $r = .65$ ($SD = 0.11$) for the balance of evidence measure, and from $r = -.43$ to $r = .25$ with a mean of $r = -.03$ ($SD = 0.18$) for the words read measure.

In principle, it would be possible to compare the individual differences and group models in a rigorous and quantitative way using Bayesian model selection (e.g., Pitt et al., 2002)

or minimum description length (e.g., Rissanen, 1996, 2001) measures, although formidable computational problems are involved in calculating the necessary definite integrals. Even a simple measure capable of controlling for model complexity, such as the Bayesian information criterion (Schwarz, 1978), requires a probabilistic measure of data fit that, in turn, would require making assumptions about confidence and response time distributions. Given the wide variety of different distributions regarded as serious theoretical contenders for response times (e.g., Luce, 1986), any such assumption would be problematic.[4] It is fortunate, therefore, that the performance of the individual and group models can be interpreted without recourse to these sorts of measures. The basic message is that there is no strong evidence of in the data of individual differences in the decision thresholds used by participants. The individual differences model is much more complicated than the group model, using $41 \times 2 = 82$ parameters rather than two, yet the average correlations with time and confidence almost all decrease. Under basic principles of model selection, this pattern of results favors the simpler account of the data that assumes there are no individual differences.

6. Application to information retrieval

Having developed and evaluated the random walk and accumulator models in relation to human performance, we examined their application to the real world text classification problem of finding documents about topics of interest in a large corpus. The Reuters-21578 corpus using the ModApte split into training and test documents is a standard information retrieval problem, for which the performance benchmarks of established machine learning techniques are available. Yang and Liu (1999) present results for five classifiers called support vector machines (SVM), k-nearest neighbor classifiers (kNN), Linear least squares fit classifiers (LLSF), neural network classifiers (NNets), and Naive Bayes classifiers (NB). SVM classifiers use the training set to solve a quadratic assignment problem that finds optimal hyperplanes separating documents into those about a topic, and those not about a topic. These hyperplanes are then applied to classify new documents from the test set. kNN classifiers establish a metric for measuring the similarity of documents, and then classify those in the test set based on the known classifications of training set documents in their proximity. LLSF classifiers generate a multivariate regression model from a training set that can be applied to new documents. The NNet classifier uses the training set to learn the connection weights for a three-layer neural network and then applies this network to classify the test set documents. The NB classifier, as described earlier, is basically a version of the random walk model that always considers every word in classifying documents.

Unfortunately, Yang and Liu's (1999) published results are based on heavily pre-processed versions of the Reuters-21578 documents, where common or "stop" words have been removed, and "stemming" algorithms have been applied in an attempt to reduce words like "fishing" and "fished" to their root word "fish." As mentioned earlier, the version of Reuters-21578 used in this study involves essentially the raw text documents, and uses a bare minimum of pre-processing. For this reason, the results obtained for the random walk and accumulator models are not directly comparable with Yang and Liu's (1999) results, although it is at least possible to use the machine learning benchmarks as rough guides to acceptable levels of performance.

To this end, both the random walk and accumulator models were applied to the Reuters-21578 problem, making classificatory decisions for each of the 3,299 test documents against each of the 90 specified topics. A large number of different parameterizations were tried for both models, using every possible combination allowed by independently choosing evidence values of 0, 5, 10, 25, 50, 100 and 200 for both the "yes" and "no" decision thresholds.

6.1. Decision performance

Two standard measures of performance, called precision and recall, are used in the information retrieval literature to measure decision accuracy for text classification. Precision measures the proportion of documents a model decides are about a topic that actually are about the topic. Recall measures the proportion of documents actually about a topic that are identified as such by the model.

Fig. 13 shows the precision and recall measures for a best performed subset of the parameterizations examined. Models are represented by markers located according to their recall (along the *x*-axis) and their precision (along the *y*-axis). Random walk models are indicated by circular makers, while accumulators are indicated by square markers, and the parameterization of the model is labeled. The precision and recall performance of the five benchmark machines learning methods are shown by filled markers.

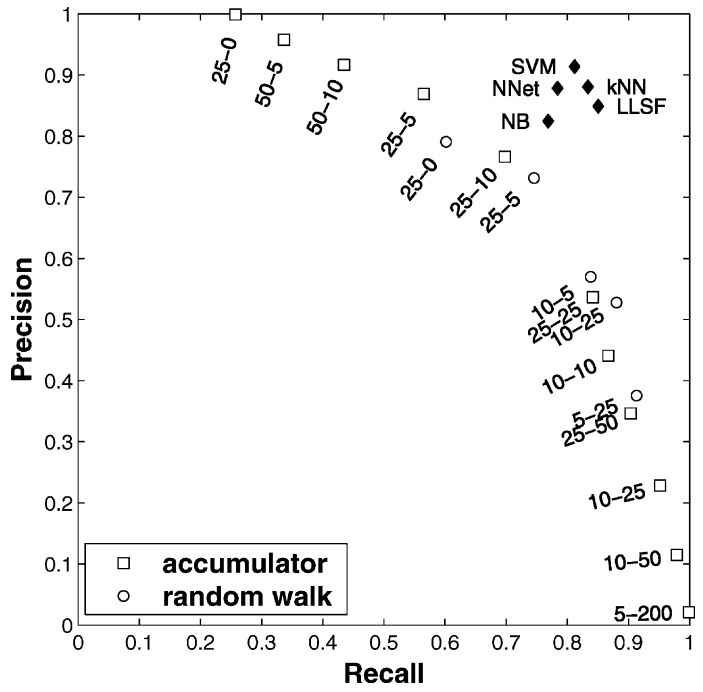

Fig. 13. The precision and recall performance of selected random walk (open circle markers) and accumulator (open square markers) models, together with the performance of established machine learning techniques (closed markers). See text for details.

Fig. 13 shows that most of the best performed models are accumulators. In particular, those models that achieve either perfect precision (with low recall), or perfect recall (with low precision) are all accumulator models. Perfect precision is achieved when a large evidence threshold of 25 is required to make "yes" decisions, but an evidence threshold of zero is used for "no" decisions. This means that the model is conservative in deciding a document is about a topic, but readily decides a document is not about a topic. Perfect recall, in contrast, is achieved when a low evidence threshold of five is used for "yes" decisions, but a very large evidence threshold of 200 is used for "no" decisions. This means that the model will decide a document is about a topic on the basis of a little evidence, but is very conservative in deciding a document is not about a topic. Parameterizations between these two extremes balance the competing demands of precision and recall according to the different levels of evidence required for "yes" and "no" decisions.

Fig. 13 shows that, across all of the parameterizations considered, neither the random walk nor accumulator model achieves the combined level of precision and recall performance produced by the machine learning techniques. There are some parameterizations, however, for which performance is competitive, particularly when viewed in an operational context. The 25–10 accumulator model and the 25–5 random walk model, for example, have almost the same recall as the machine learning benchmarks, with a 10–20% difference in precision. In practice, this means that the sequential sampling models find essentially the same number of documents that are about topics, but return an extra one or two false alarms for every set of 10 documents found.

In any case, as noted above, some of the observed deficiency in our results can be attributed directly to the pre-processing of the corpus. To gauge the magnitude of this difference, we applied the NB classifier to our version of the corpus, and measured its precision to be 0.55 with a recall of 0.86. Given Yang and Liu's (1999) results for the NB classifier, with a recall of 0.77 and precision of 0.82, it seems clear that the lack of pre-processing does penalize the random walk and accumulator results reported here. For this reason, a comparison of established machine learning techniques with sequential sampling models on exactly the same problem is a worthwhile topic for further research.

On a different front, it is interesting to note that Fig. 13 approximately represents the results of fitting the random walk and accumulator models to the human judges who "ground-truthed" the Reuters-21578 corpus. A complete account of this human decision making process would result in a model achieving perfect precision and recall. In this sense, to the extent that the random walk and accumulator models achieve their best combined precision and recall performance under the parameterizations 25–5 and 25–10, respectively, these parameterizations constitute their best fit to the human data.

The asymmetry in the best parameterizations is different from that obtained from our experimental data, with more evidence being required to make a "yes" than a "no" decision. This difference seems to be explained naturally in terms of the different task demands involved. For the participants in our experiment, relatively few decisions needed to be made, and the expected base rate of "yes" to "no" decision was presumably about fifty-fifty. Under these conditions, it seems reasonable to be more comfortable deciding a document is about a topic, and more cautious about deciding (before all of the document has been seen) that it is not about a topic. The humans who evaluated the Reuters-21578 corpus, however, had to

evaluate 21,578 documents against 90 topics, knowing that very few of the document–topic combinations would require a "yes" decision, but that it was important these combinations were found. Under these conditions, it seems reasonable to be conservative in making "yes" decisions, to avoid false alarms, but to require only moderate evidence before making a "no" decision, to quicken the decision making process. These sorts of utilities in decision making are naturally handled within Bayesian decision theory (e.g., Lindley, 1972, pp. 1–3), and so it should be possible to extend sequential sampling process models to give a rational account of the different thresholds used in different contexts.

6.2. Time performance

Consistent with the overwhelming focus placed on decision accuracy by the information retrieval literature, Yang and Liu (1999) do not provide measures of the time taken by the various machine learning techniques to make their decisions. Timeliness is clearly, however, an important determinant of relative performance in an applied setting. In general, the machine learning techniques involve considerable levels of computation, either during the training process, the process of classifying new documents, or both. The quadratic assignment problem solved by SVMs, for example, become computationally intensive for very large problems, LLSF classifiers must solve a large least-squares problem, and NNets are notoriously time consuming to train. When classifying new documents, most existing machine learning techniques consider every word in the document, and often have to calculate involved functions. It would, of course, be possible to apply any of the machine learning algorithms to a limited number of words at the beginning of each document. This will clearly improve their time performance, but how precision and recall are affected is an open (and interesting) empirical question.

The random walk and accumulator models, in contrast, are exceptionally easy to train and make very fast decisions because they are explicitly designed to read only as many words in the document as they require. Training involves only calculating the evidence value for each word in relation to each topic using Eq. (1), which can be achieved by examining each word in the training set of documents exactly once, and maintaining counts of how often a word is seen in documents about and not about a topic. Classifying new documents involves looking up appropriate evidence values and adding them to a counter or counters until a pre-specified threshold total is reached. These are very simple computational processes that can be implemented efficiently in software.

In terms of time performance, however, the greatest strength of the sequential sampling models comes from the limited number of words that need to be read to make a decision. Table 3 shows the mean number of words read by the best performed models detailed in Fig. 13, and the minimum and maximum number of words read. For almost all of the parameterizations, decisions are made using a remarkably small number of words on average, given that the mean number of words in the test documents is 121. The 25–5 random walk model and the 25–10 accumulator model, for example, examine an average of fewer than five and six words per document, respectively, which corresponds to less than 5% and 4% of the mean document length. In practical terms, this means that these models are able to make decisions almost as accurately as machine learning benchmarks after examining less than 5% of the data, and so provide comparable decision performance significantly more quickly. On the basis of these

Table 3
Mean number of words examined, and the range, for various random walk and accumulator models

Model type	Thresholds	Mean words	Range
Accumulator	25–0	1.4	1–30
Accumulator	50–5	3.4	1–102
Accumulator	50–10	6.0	1–114
Accumulator	25–5	3.3	1–80
Random walk	25–0	2.0	1–292
Accumulator	25–10	5.8	1–83
Random walk	25–5	4.5	1–666
Random walk	10–5	3.8	1–297
Accumulator	25–25	12.8	3–147
Random walk	10–25	11.9	2–97
Accumulator	10–10	5.5	1–64
Random walk	5–25	14.0	1–441
Accumulator	25–50	24.0	3–186
Accumulator	10–25	11.9	2–97
Accumulator	10–50	21.6	2–140
Accumulator	5–200	34.7	1–207

The mean number of words in the test documents is 121. The ordering of the different parameterizations corresponds to Fig. 13.

results, it seems reasonable to assert that the random walk and accumulator models are superior to established machine learning techniques on any sensible "decision performance per unit computation" measure.

It is worth emphasizing that the speed advantages of the random walk and accumulator models follow directly from the psychological observations on which they are based. The environmental regularity that words near the beginning of a document will be the most useful allows for fast and accurate decisions. In terms of speed, it is particularly important that non-compensatory "no" decisions are made, since most of the documents in Reuters-21578 are not about most of the topics. It is the idea that different decisions are effectively competing explanations of observed data that achieves this, because evidence can be accrued directly in favor of a "no" decision. Many established text classifiers based on machine learning algorithms do not operate this way. Instead, they construct a measure of the similarity between the document in question, and some abstract representation of the topic in question. When the measure of similarity exceeds some criterion value, the decision is made that the document is about the topic, otherwise the default decision is made that the document is not about the topic. This means that, in principle, these classifiers must examine all of the words in a document before making a "no" decision.

6.3. Using confidence to prioritize

The third psychological observation, that the classification decision making process generates a number of related performance measures, also has applied benefits. In an information retrieval context the confidence ratings can be used to prioritize the documents returned by a

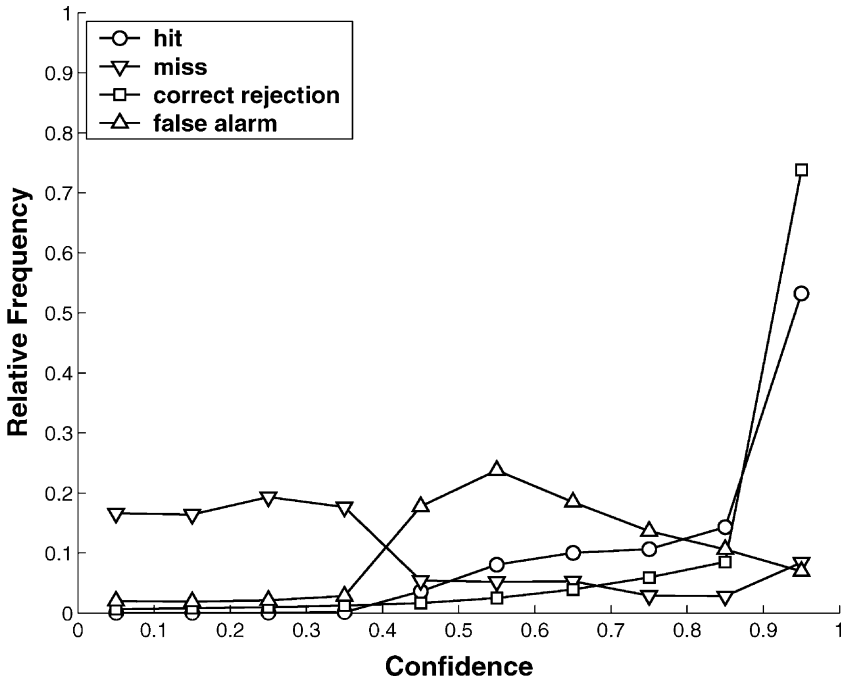

Fig. 14. Confidence distributions for the 25–10 forced choice accumulator model.

search, through a familiar mechanism such as a "relevancy score." The obvious way to perform prioritization is to order all of the document–topic combinations, starting with those classified as about the topic with highest confidence down to those with the lowest confidence, and then appending those document–topic combinations classified as not about the topic, starting with the lowest confidence and continuing to the highest confidence decision.

This prioritization scheme will be effective to the extent that confidence measures provide an accurate assessment of the outcome of the decision making process, having high confidence when the decision is correct and low confidence when it is not. Fig. 14 presents an analysis of the ability of the 25–10 accumulator model to do this using the balance of evidence confidence measure given by Eq. (3). The confidence distributions across all decisions are shown for the four signal detection classes of hit, miss, correct rejection and false alarm. These distributions are meaningful, in the sense that the model often has high confidence when it makes a hit or correct rejection decision, and generally has lower confidence when it produces a miss or a false alarm.

Fig. 15 presents the same analysis for the 25–5 random walk model using the confidence measure determined by the number of words read. These confidence distributions are far less meaningful, and display some serious problems. For example, the model more often has high confidence when it misses than when it hits.

The impact on prioritization of these differences in confidence measures is summarized in Fig. 16. This "effort–reward" graph shows the proportion of relevant documents (i.e., the reward) found by working through a given proportion of the prioritized list (i.e., the effort) for

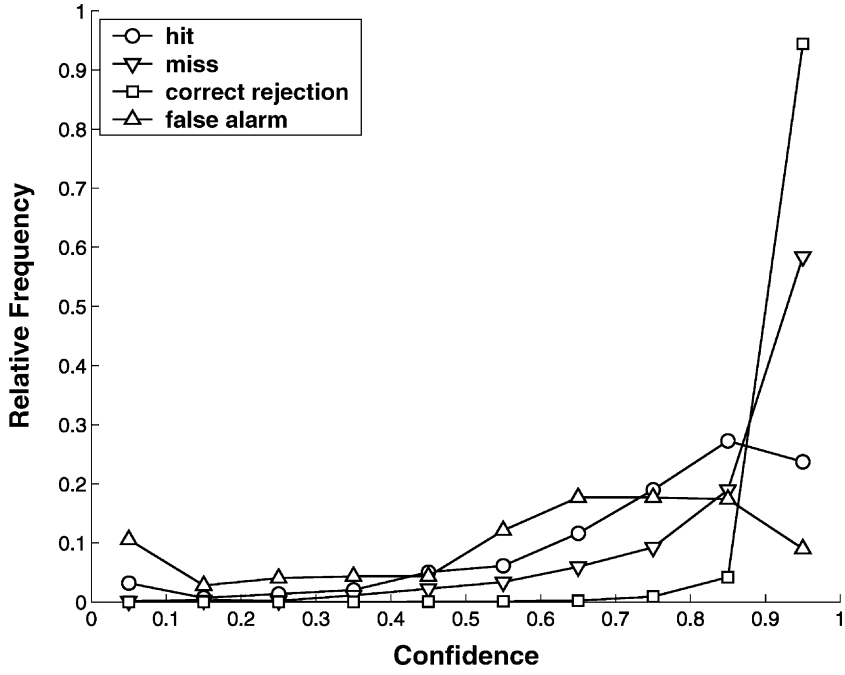

Fig. 15. Confidence distributions for the 25–5 forced choice random walk model.

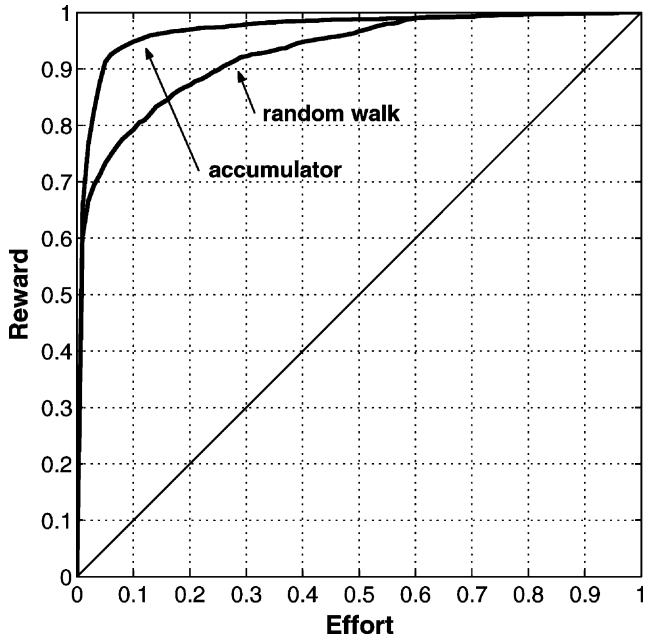

Fig. 16. Effort–reward performance for prioritization using the accumulator and random walk models.

both the random walk and accumulator models. Both approaches result in about 65% of the relevant document–topic combinations being placed in the first 1% of the list, but the accumulator then performs significantly better, allowing about 95% of the relevant document–topic combinations to be found by examining the top 10% of the list. In this way, the theoretical benefits of the balance of evidence measure found in modeling human data also have applied benefits in allowing more effective prioritization.

7. General discussion

There are two main avenues for improving sequential sampling accounts of text classification as models of human performance, and in their application to information retrieval. The first possibility does not involve changing the sequential sampling processes themselves, but rather seeks to provide them with better information about a document than is currently provided by the evidence values. The second possibility does involve changing the sequential sampling processes, by extending them to have more realistic and useful memory and learning capabilities.

7.1. Alternative evidence measures

The evidence values defined in Eq. (1) basically measure, on a log-odds scale, the frequency with which a word occurs in topic versus non-topic documents. While this approach has the advantage of being easily interpreted and calculated, it is an overly simple model of human semantic representation. When people read text documents, meaning can be accrued from parts of words, or sequences of words, in a way that is mediated by the earlier content of the document. Providing sequential sampling models with evidence values that captured some of this richness in meaning seems likely to improve their performance, without requiring additional complexity in the information accrual process. In particular, the currently implausible independence assumption in Eq. (2), where the evidence provided by each word is assumed to be constant regardless of surrounding words, would become more plausible if appropriate sequences of words, such as key phrases, were the basic unit of evidence.

A variety of more sophisticated evidence measures have been developed in the psychological and information retrieval literatures. These include the representational vectors generated by the latent semantic analysis (Landauer & Dumais, 1997) and hyperspace analogue to language (Lund & Burgess, 1996) approaches, as well as the probabilistic method developed by Griffths and Steyvers (2002), all of which measure, in various ways, the patterns with which words appear in the same and different contexts. At the other end of the spectrum, the *n*-gram approach (Damashek, 1995) uses sequences of successive characters as its basic representational unit, and has been shown to capture a surprising level of semantic information. How alternative measures of evidence such as these affect the performance of random walk and accumulator models is an interesting question for future research.

7.2. Incorporating memory and learning

Whatever form of the semantic information they receive, both the random walk and accumulator models considered here lack two basic psychological ingredients as models of human de-

cision making: they do not involve any form of memory, nor do they adapt to their environment in any self-regulating way. The evidence values of words are accrued with perfect accuracy, without forgetting or distortion over time or as new information arrives, and the thresholds that determine decision making do not adapt automatically to changes in task demands or the nature of the stimulus environment. Clearly, real world human text classification involves a limited memory and a potential to learn and adapt, and so the extension of the models to incorporate these characteristics is important theoretically. Previous modeling of human decision making using sequential sampling processes has considered both of these issues extensively (e.g., Pietsch & Vickers, 1997; Ratcliff, 1987; Smith, 2000; Vickers, 1985), and so candidate extended models should be easy to develop. The main challenge involves the collection of human performance data that allow the relative merits of these models to be assessed. This might involve considering the classification decisions made by people on very long documents, where memory is likely to be important, or presenting document sets with very different "yes" and "no" base rates, where the adaptation of decisions thresholds is likely to be important.

8. Conclusion

In Section 1, we argued that the decision processes involved in human text document classification are interesting from both theoretical and applied perspectives. Theoretically, text documents provide a ready source of richly structured real world stimuli, and so force quantitative accounts of human decision making to consider the role of the environment in relation to internal cognitive processes. In terms of applications, the ability to classify text documents automatically is a central problem in information retrieval. We conclude by drawing some implications of the results presented in this paper for both the theoretical and applied problems.

On the applied front, the random walk and accumulator models have been shown to make classificatory decisions that are competitive with benchmark machine learning techniques on a standard problem, and are able to make these decisions much more quickly. In any applied setting where timeliness competes with accuracy as a criterion for good performance, there are grounds for regarding the sequential sampling models as superior. In addition, the ability of the accumulator model to produce sensible confidence measures has the applied advantage of allowing the results of its decision making to be prioritized.

On the theoretical front, the ability of the accumulator model to predict classificatory decisions that the random walk cannot suggests that it is a superior sequential sampling process account of human decision making. Further impetus for accepting the accumulator model comes from its ability to correlate well with the change in the number of words read by participants across speed and accuracy conditions through sensible and interpretable adjustments in its evidence threshold parameters. Finally, the balance of evidence measure of confidence, which is only possible under the accumulator approach to information accrual, was also found to correlate well with human performance.

It is interesting to note that the observed differences between the random walk and accumulator models arise, at least in part, from the non-stationary evidence structure of the text document stimulus domain (recall, in particular, Fig. 12). The implication is that developing and distinguishing between cognitive models can be advanced by using natural environmental

stimuli. The text classification problem is particularly well suited to studying the ecological rationality of non-compensatory decision making in structured environments, because documents have such an obvious sequential information structure. We would argue, however, that any decision making problem where the search for relevant information is effortful, but can productively follow a non-arbitrary pattern of search, should support non-compensatory decision making. For example, diners outside an unfamiliar restaurant can decide whether or not it is vegetarian from the first four or five main courses, and do not need to read the entire menu. In many football codes, an experienced fan can determine the coach of a team by scanning the playing field in an ordered way, and does not need to locate every player to make a decision. In this way, understanding the information structure of environments provides an opportunity to understand the rational basis and processing strategies of human decision making. The general lesson, we believe, is that real world decision tasks with richly structured stimuli can productively be applied in developing and evaluating models of human cognitive processes.

Notes

1. We thank Josh Tenenbaum for suggesting this interpretation.
2. We do not report standard null-hypothesis significance testing (NHST) inferences, because we are sensitive to criticisms of this approach (e.g., Cohen, 1994; Edwards, Lindman, & Savage, 1963; Howson & Urbach, 1993; Hunter, 1997; Lindley, 1972) including, in particular, that NHST violates the likelihood principle, and so does not satisfy a basic requirement for rational, consistent and coherent statistical decision making.
3. This is especially true given the evidence structure of the document corpus observed in Fig. 1 since, on average, the absolute evidence value decreases as words in a document are read.
4. It is true that the text classification task generally takes longer than most tasks for which response time distributions are studied, although with mean response times of 18.0 and 32.9 s for the speed and accuracy conditions, respectively, it is comparable to some expanded judgment tasks (e.g., Pietsch & Vickers, 1997).

Acknowledgments

This research was supported by the Australian Defence Science and Technology Organisation. We wish to thank Helen Braithwaite, Peter Bruza, Nick Burns, Marcus Butavicius, Lama Chandrasena, Simon Dennis, Brandon Pincombe, Douglas Vickers, and Chris Woodruff for helpful comments, and John Anderson, Josh Tenenbaum, and two anonymous reviewers for helpful comments on an earlier version of this paper.

References

Anderson, J. R. (1990). *The adaptive character of though*. Hillsdale, NJ: Lawrence Erlbaum.
Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Anderson, J. R. (1992). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–517.
Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
Bourne, L. E. (1974). An inference model of conceptual rule learning. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 231–256). Potomac, MD: Lawrence Erlbaum.
Brunswik, E. (1943). Organismic achievement and environmental probabilities. *Psychological Review*, *50*, 255–272.
Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, *32*(2), 91–134.
Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). New York: Chapman & Hall.
Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.
Damashek, M. (1995). Gauging similarity with *n*-grams: Language-independent categorization of text. *Science*, *267*, 843–848.
Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242.
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.
Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
Griffths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th annual conference of the cognitive science society* (pp. 381–386).
Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach*. La Salle, IL: Open Court Press.
Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*(1), 3–7.
Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104*, 344–366.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.
Laming, D. R. J. (1968). *Information theory and choice-reaction time*. London: Academic Press.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.
Leonard, T., & Hsu, J. S. J. (1999). *Bayesian methods: An analysis for statisticians and interdisciplinary researchers*. New York: Cambridge University Press.
Lewis, D. D. (1997). *Reuters-21578 text categorization test collection*. Available at http://www.research.att.com/~lewis/reuters21578/readme.txt.
Lindley, D. V. (1972). *Bayesian statistics: A review*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*, 77–105.
Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.
Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments, & Computers*, *28*(2), 203–208.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification. *Psychological Review*, *85*, 207–238.
Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.
Myung, I. J., & Shepard, R. N. (1996). Maximum entropy inference and stimulus generalization. *Journal of Mathematical Psychology*, *40*, 342–347.
Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*(2/3), 103–134.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.
Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300.
Pietsch, A., & Vickers, D. (1997). Memory capacity and intelligence: Novel techniques for evaluating rival models of a fundamental information processing mechanism. *Journal of General Psychology*, *124*, 229–339.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R. (1987). More on the speed and accuracy of positive and negative responses. *Psychological Review*, *88*, 552–572.

Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*(1), 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*(5), 1712–1717.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin* & *Review*, *1*(1), 2–28.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classification. *Psychological Monographs*, *75*(13), Whole No. 517.

Simon, H. A. (1956). Rational choice and the structure of environments. *Psychological Review*, *63*, 129–138.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, *44*, 408–463.

Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*, 727–780.

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two reaction-time models applied to perceptual matching. *Psychonomic Bulletin* & *Review*, *7*, 208–256.

Vickers, D. (1979). *Decision processes in visual perception*. New York, NY: Academic Press.

Vickers, D. (1985). Antagonistic influences on performance change in detection and discrimination tasks. In G. d'Ydewalle (Ed.), *Cognition, information processing and motivation, Amsterdam* (pp. 79–115). North-Holland.

Vickers, D. (2001a, September). *The elusive "interval of uncertainty:" Evidence for a dynamic normalizing of sensory magnitudes*. Paper presented at the 32nd meeting of the European mathematical psychology group, Lisbon, Portugal.

Vickers, D. (2001b). Where does the balance of evidence lie with respect to confidence? In E. Sommerfeld, R. Kompass, & T. Lachmann (Eds.), *Proceedings of the seventeenth annual meeting of the international society for psychophysics* (pp. 148–153). Lengerich: Pabst.

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments. I. Properties of a self-regulating accumulator module. *Non-linear Dynamics, Psychology, and Life Sciences*, *2*(3), 169–194.

Vickers, D., & Pietsch, A. (2001). Decision-making and memory: A critique of Juslin and Olsson's (1997) sampling model of sensory discrimination. *Psychological Review*, *108*(4), 789–804.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 42–49). Berkley, CA: ACM.