A Prioritization Model for Suicidality Risk Assessment

Han-Chin Shing
Computer Science
University of Maryland
College Park, MD
shing@cs.umd.edu

Philip Resnik
Linguistic/UMIACS
University of Maryland
College Park, MD
resnik@umd.edu

Douglas W. Oard
iSchool/UMIACS
University of Maryland
College Park, MD
oard@umd.edu

Abstract

We reframe suicide risk assessment from social media as a ranking problem whose goal is maximizing detection of severely at-risk individuals given the time available. Building on measures developed for resource-bounded document retrieval, we introduce a well founded evaluation paradigm, and demonstrate using an expert-annotated test collection that meaningful improvements over plausible cascade model baselines can be achieved using an approach that jointly ranks individuals and their social media posts.

1 Introduction

Mental illness is one of the most significant problems in healthcare: in economic terms alone, by 2030 mental illness worldwide is projected to cost more than cardiovascular disease, and more than cancer, chronic respiratory diseases, and diabetes combined (Bloom et al., 2012). Suicide takes a terrible toll: in 2016 it became the second leading cause of death in the U.S. among those aged 10-34, fourth among those aged 35-54 (Hedegaard et al., 2018). Prevalence statistics suggest that roughly 141 of the 3,283 people who attended ACL 2019 have since had serious thoughts of suicide, 42 have made a plan, and 19 have actually made attempts.[1]

The good news is that NLP and machine learning are showing strong promise for impact in mental health, just as they are having large impacts everywhere else. Traditional methods for predicting suicidal thoughts and behaviors have failed to make progress for fifty years (Franklin et al., 2017), but with the advent of machine learning approaches (Linthicum et al., 2019), including text analysis methods for psychology (Chung and Pennebaker, 2007) and the rise of research on mental health using social media (Choudhury, 2013), algorithmic classification has reached the point where it can now dramatically outstrip performance of prior, more traditional prediction methods (Linthicum et al., 2019; Coppersmith et al., 2018). Further progress is on the way as the community shows increasing awareness and enthusiasm in this problem space (e.g., Milne et al., 2016; Losada et al., 2020; Zirikly et al., 2019).

The bad news is that moving these methods from the lab into practice will create a major new challenge: identifying larger numbers of people who may require clinical assessment and intervention will increase stress on a severely resource-limited mental health ecosystem that cannot easily scale up.[2] This motivates a reformulation of the technological problem from classification to *prioritization* of individuals who might be at risk, for clinicians or other suitably trained staff as downstream users.

Perhaps the most basic way to do prioritization is with a single priority queue that the user scans from top to bottom. This "ranked retrieval" paradigm is common for Information Retrieval (IR) tasks such as document retrieval. The same approach has been applied to ranking people based on their expertise (Balog et al., 2012), or more generally to ranking entities based on their characteristics (Balog, 2018). Rather than evaluating categorical accuracy, ranked retrieval systems are typically evaluated by some measure of search quality that rewards placing desired items closer to the top (Voorhees, 2001). Most such measures use only item position, but we find it important to also model the *time* it takes to recognize desired items, since in our setting the time of qualified users is the most limited resource.

In this paper, we do so by building on Time-

[1] Approximately: ACL is international, but these figures use prevalence statistics for U.S. adults (SAMHSA, 2019).

[2] 120M Americans live in areas with mental healthcare provider shortages (Bureau of Health Workforce, 2020). That number reflects an increase of about 7 million people between September 30, 2019 and March 31, 2020.

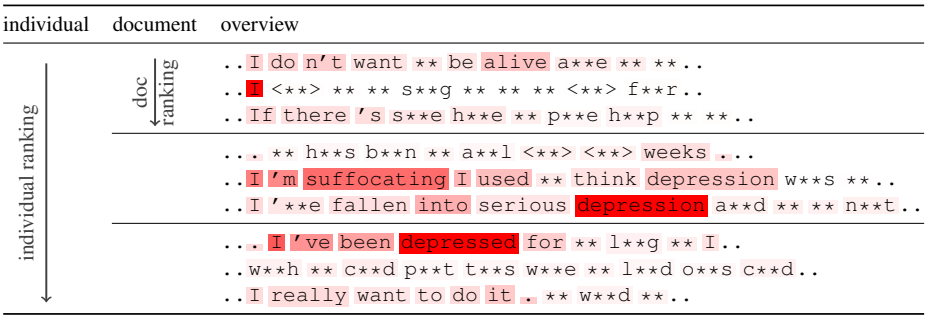


Figure 1: Illustration of an assessment framework in which individuals are ranked by predicted suicide risk based on social media posts, posts are ranked by expected usefulness for downstream review by a clinician, and word-attention highlighting helps foreground important information for risk assessment. Real Reddit posts, obfuscated and altered for privacy.

Biased Gain (TBG, Smucker and Clarke, 2012), an IR evaluation measure that models the expected number of relevant items a user can find in a ranked list given a time budget. We observe that in many risk assessment settings (e.g., Yates et al. (2017); Coppersmith et al. (2018); Zirikly et al. (2019)), the available information comprises a (possibly large and/or longitudinal) set of documents, e.g. social media posts, associated with each individual, of which possibly only a small number contain a relevant signal.[3] This gives rise to a formulation of our scenario as a nested, or *hierarchical*, ranking problem, in which individuals are ordered by priority, but each individual's documents must also be ranked (Figure 1). Accordingly, we introduce hierarchical Time-Biased Gain (hTBG), a variant of TBG in which individuals are the top level ranked items, and expected reading time is modeled for the ranked list of documents that provides evidence for each individual's assessment. In addition, we introduce a prioritization model that uses a three-level hierarchical attention network to jointly optimize the nested ranking task; this model also addresses the fact that in our scenario, as in many other healthcare-related scenarios, relevance obtains at the level of individuals rather than individual documents (cf. Shing et al., 2019). Using a test collection of Reddit-posting individuals who have been assessed for suicide risk by clinicians based on their posts (Shing et al., 2018), we use hTBG to model prioritization of individuals and demonstrate that our joint model substantially outperforms cascade model baselines in which the nested rankings are produced independently.

[3]Our dataset, for example, has one severe risk individual with 1,326 postings, of which only two are "signal" posts identified by the experts. See Table 2 for detailed statistics.

2 Related Work

NLP for Risk Assessment. Calvo et al. (2017) survey NLP for mental health applications using non-clinical texts such as social media. Several recent studies and shared tasks focus on risk assessment of individuals in social media using a multi-level scale (Milne et al., 2016; Yates et al., 2017; Losada et al., 2020). Shing et al. (2018) introduce the dataset we use, and Zirikly et al. (2019) describe a shared task in which 11 teams tackled the individual-level classification that feeds into our prioritization model (their Task B). Our work contributes by modeling the downstream users' prioritization task as taking a key step closer to the real-world problem.

Hierarchical Attention Attention, especially in the context of NLP, has two main advantages: it allows the network to attend to likely-relevant parts of the input (either words or sentences), often leading to improved performance, and it provides insight into which parts of the input are being used to make the prediction. These characteristics have made attention mechanisms a popular choice for deep learning that requires human investigation, such as automatic clinical coding (Baumel et al., 2018; Mullenbach et al., 2018; Shing et al., 2019). Although concerns about using attention for interpretation exist (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Wallace, 2019), Shing et al. (2019) show hierarchical document attention can align well with human-provided ground truth.

Our prediction model, 3HAN, is a variant of Hierarchical Attention Networks (HAN, Yang et al., 2016). Yang et al. use a two-level attention mechanism that learns to pay attention to specific words in a sentence to form a sentence representation, and at the next higher level to weight specific sentences in

a document in forming a document representation. Adapting this approach to suicide assessment of at-risk individuals, our model moves a level up the representational hierarchy, learning also to weight documents to form representations of individuals. This allows us to jointly model ranking individuals *and* ranking their documents as potentially relevant evidence, without document-level annotations.

Evaluating rankings. There is an extensive IR literature on quality measures for ranked lists (Järvelin and Kekäläinen, 2002; Chapelle et al., 2009; Smucker and Clarke, 2012; Sakai, 2019), which generally reward placing highly relevant items near the top of the list, and are often relatively insensitive to mistakes made near the bottom.

In the setting of suicidality risk assessment, we care about how much gain (number of at-risk individuals found) can be achieved for a given time budget. Time-biased gain (TBG, Smucker and Clarke, 2012) measures this by assuming a determined user working down a ranked list, with the discount being a function of the time it takes to reach that position. However, neither TBG nor other ranking measures, to the best of our knowledge, can measure the *hierarchical* ranking found in the scenario that motivates our work: ranking items (i.e. individuals) when each item itself contains a ranked list of potential evidence (their posts). In this paper, we design a new metric, hierarchical time-biased gain (hTBG), to measure the hierarchical ranking by incorporating the cascading user model found in Expected Reciprocal Rank (ERR, Chapelle et al., 2009) into TBG.

3 A Measure for Risk Prioritization

Section 1 argued for formulating risk assessment as a prioritization process where the assessor has a limited time budget. This leads to four desired properties in an evaluation measure:[4]

- **Risk-based**: Individuals with high risk should be ranked above others.
- **Head-weighted**: Ranking quality near the top of the list, where assessors are more likely to assess, should matter more than near the bottom.
- **Speed-biased**: For equally at-risk individuals, the measure should reward ranking the one who can be assessed more quickly closer to the top, so that more people at risk can be identified within a given time budget.
- **Interpretable**: The evaluation score assigned to a system should be meaningful to assessors.

[4]Throughout, *assessor* or *user* signify a clinician or other human assessor, and *individual* is someone being assessed.

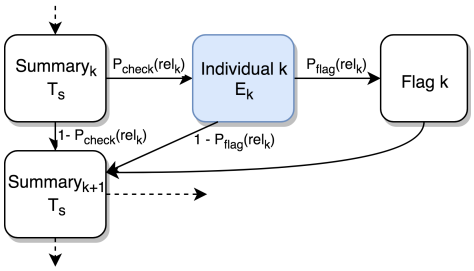


Figure 2: User model for Time-Biased Gain (TBG)

Among many rank-based measures that satisfy the *risk-based* and *head-weighted* criteria, TBG directly accounts for assessment time in a way that also satisfies the *speed-biased* criterion (see Theorem 3.1). Furthermore, the numeric value of TBG is a lower bound on the expected number of relevant items — in our case, high-risk individuals — found in a given time budget (Smucker and Clarke, 2012), making it *interpretable*. After introducing TBG, in Section 3.2 we develop *hierarchical* Time-Biased Gain (hTBG), an extension of TBG, to account for specific properties of risk assessment using social media posts.[5]

3.1 Time-Biased Gain

TBG was originally developed in IR for the case of a user seeking to find a relevant document, but here we frame it in the context of risk assessment (Figure 2). TBG assumes a determined user (say a clinician) examining a ranked list of individuals in the order presented by the system. For each individual, the clinician first examines a *summary* and then decides whether to check relevance via more detailed examination, or to move on. Checking requires more time to make an assessment of whether the individual is indeed at-risk. TBG is a weighted sum of gain, g_k, and discount, $D(\cdot)$, a function of time:

$$\text{TBG} = \sum_{k=1}^{\infty} g_k D\left(T\left(k\right)\right). \qquad (1)$$

[5]TBG and hTBG code: https://github.com/sidenver/hTBG

Parameter	Description	Value
$P_{\text{check}}(\text{rel}_i)$	Prob. to check, given the relevance of summary	0.64, if $\text{rel}_i = 1$ 0.39, if $\text{rel}_i = 0$
$P_{\text{flag}}(\text{rel}_i)$	Prob. to flag, given the relevance of individual	0.77, if $\text{rel}_i = 1$ 0.27, if $\text{rel}_i = 0$
T_s	Seconds to evaluate a summary	4.4
$T_\alpha W + T_\beta$	Seconds to judge W words	$0.018W + 7.8$

Table 1: Parameters used for TBG and hierarchical TBG.

$T(k)$ is the expected amount of time it takes a user to reach position k:

$$T(k) = \sum_{i=1}^{k-1} t\,(i) \tag{2}$$

$$t(i) = T_s + P_{\text{check}}\,(\text{rel}_i)\,E_i \tag{3}$$

where $t(i)$ is expected time spent at position i. Breaking down $t(i)$, T_s is the time it takes to read a summary and decide whether to check the individual; if yes (probability $P_{\text{check}}(\text{rel}_i)$), E_i is expected time for detailed assessment, calculated as a function of the individual's total word count W_i:

$$E_i = T_\alpha W_i + T_\beta \tag{4}$$

where T_α and T_β scales words to time. The discount function $D(t)$ decays exponentially with half-life h:

$$D(t) = 2^{-\frac{t}{h}} \tag{5}$$

where h is the time at which half of the clinicians will stop, on average. The expected stop time (or mean-life) is $\frac{h}{ln(2)}$. Finally, the gain, g_k is:

$$g_k = P_{\text{check}}(\text{rel}_k) P_{\text{flag}}(\text{rel}_k) \mathbb{1}_{[rel_k=1]} \tag{6}$$

where $P_{\text{check}}(\text{rel}_k)$ is the probability of checking the individual after reading the summary at position k, and $P_{\text{flag}}(\text{rel}_k)$ is the probability of then flagging that individual as high risk. Gain thus accrues only if a clinician actually finds a high-risk individual.

The decay function in Equation 5 monotonically decreases with increasing time (and thus rank), so TBG satisfies the *head-weighted* criterion. Table 1 shows the parameters used in Smucker and Clarke (2012), which were estimated from user studies using data from TREC 2005 Robust track.

Particularly of interest in a time-limited assessment, we can prove that TBG is *speed-biased*:

Theorem 3.1 (TGB satisfies the speed-biased criterion)**.** *Swapping an at-risk individual of longer assessment time ranked at k with an equally at-risk individual of shorter assessment time ranked at $k + r$, where $r > 0$, always increases TBG.*

Proof. See Appendix B.1 □

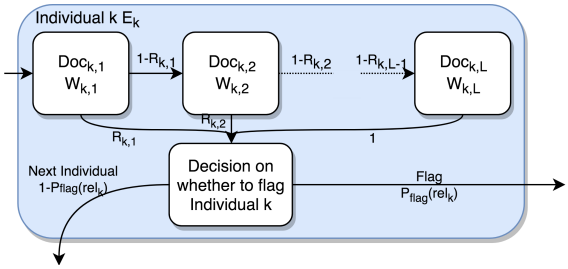


Figure 3: hTBG's model for calculating expected assessment time for an individual, replacing shaded box in Figure 2.

3.2 Hierarchical Time-Biased Gain

TBG assumes that detailed assessment involves looking at *all* available evidence (Equation 4). However, in our setting, an individual may have a large or even overwhelming number of social media posts. One severe risk individual in the SuicideWatch dataset, for example, has 1,326 posts in Reddit, the vast majority of which would provide the assessor with no useful information. Therefore we need to prioritize the documents to be read, and a way of estimating when the user will have read enough to make a decision.

In general, clinicians engage in a sensemaking process as they examine evidence, and modeling the full complexity of that process would be difficult. We therefore make two simplifying assumptions: (1) that there is a high-signal document that suffices, once read, to support a positive relevance judgment, and (2) that the clinician will not read more than some maximum number of documents. These assumptions align well with those of Expected Reciprocal Rank (ERR), whose cascading user model assumes that as the user works down a ranked list (in our case, the ranked documents posted by a single individual), they are more likely to stop after viewing a highly relevant document than after viewing an irrelevant one, as their information need is more likely to have been satisfied (Chapelle et al., 2009). This results in a cascade model of user behavior: ERR $= \sum_{k=1}^{\infty} \frac{1}{k} P\,(\text{stop at k})$, in which $P\,(\text{stop at k}) = R_k \prod_{i=1}^{k-1} (1 - R_i)$, where $R_k = f(rel_k)$ is the probability of stopping at position k as a function of relevance.

This suggests replacing Equation 4 with the following expected time estimate for detailed assessment of an individual:

$$E_i = T_\alpha \sum_{l=1}^{L} \left(W_{i,l} \prod_{m=1}^{l-1} (1 - R_{i,m}) \right) + T_\beta \quad (7)$$

where $R_{i,l}$ is the probability of stopping at the l-th document for individual i, and $W_{i,l} > 0$ is the cost (in our case, word count) of reading the l-th document for individual i. Note that for the special case of $\forall i, l \in N, R_{i,l} = 0$, hTBG reduces to TBG. See Figure 3 for an illustration of E_i of hTBG. For derivation of Equation 7 from ERR's cascading user model, see Appendix B.3.

3.3 Optimal Values for TBG and hTBG

Calculation of the optimal value for a measure is often important for normalization, though not always easy; in some cases it can be NP-hard (Agrawal et al., 2009, ERR-IA). Another popular approach is to normalize by calculating the metric with an ideal collection. For example, Smucker and Clarke (2012) calculate the normalization factor of TBG by assuming a collection with an infinite number of relevant documents, each of which lack any content. In our case, however, we are actually interested in an optimal value achievable for a given test collection: the optimal values of TBG and hTBG are properties of the bottleneck that occurs due to the user's limited time-budget. We find that:

Theorem 3.2 (Optimal TBG)**.** *The optimal value of TBG under binary relevance is obtained if and only if (1) all at-risk individuals are ranked above not-at-risk individuals, and (2) within the at-risk individuals, they are sorted based on time spent in ascending order.*

Proof. See Appendix B.1 □

Theorem 3.2 makes sense, as any time spent on assessing a not-at-risk individual is time not spent on assessing other potentially at-risk individuals. Preference in assessing individuals with shorter assessment time also increased the chance of assessing more individuals in the given time budget.

Minimum Individual Assessment Time. To calculate optimal hTBG, we need to minimize individual assessment time. A natural question to ask, then, is whether a result similar to Theorem 3.2 holds for the individual assessment time of hTBG in Equation 7. By swapping paired documents, we can use proof by contradiction to show that:

Theorem 3.3. *Minimum individual assessment time is obtained if the documents are sorted in descending order by* $\frac{R_{i,l}}{W_{i,l}}$.

Proof. See Appendix B.2 □

Theorem 3.3 shows a surprisingly intuitive tradeoff between how relevant a document might be, and how much time (proportional to word counts) the expert needs to take to read it: highly relevant documents with short reading time are preferred.

Observe that Theorem 3.1 (speed-biased criterion) and Theorem 3.2 both apply to hTBG, as the two theorems only concern the ranking of individuals, not documents, and hTBG is an extension of TBG to measure the document ranking. Using Theorem 3.3 and Theorem 3.2, calculation of optimal TBG and hTBG values is simply a matter of sorting. For TBG, time complexity is $O(n \log(n))$, where $n \leq K$ is the number of at-risk individuals in the test collection. For hTBG, worst-case time complexity is $O(n \log(n) + nm \log(m))$, where $m \leq L$ is the maximum number of relevant documents per individual.

4 Classification Model

We began by motivating risk assessment via social media as a person-centered, time-limited prioritization problem, in which the technological goal is to support downstream clinicians or other assessors in identifying as many people at risk as possible. This led to the conclusion that systems should not only rank individuals but, for each individual, rank their posts, and we introduced an evaluation framework that involves an abstraction of the user's process of identifying people at risk given a nested ranking.

Next, we need a system that can produce such nested rankings of individuals and their posts. Ideally such a system should be able to train on only individual-level, not document-level, labels, since suicide risk is a property of individuals, not documents, and document labels are more difficult to obtain. In addition, such a system should ideally produce additional information to help the downstream user — if not justification of its output, then at least highlighting potentially useful information.

To address this need, we introduce 3HAN, a hierarchical attention network (Yang et al., 2016) that extends up to the level of individuals, who are

represented as sequences of documents. This architecture is similar to the network we proposed in Shing et al. (2019) for coding clinical encounters; it obtained good predictive performance and we also showed that, despite concerns about the interpretation of network attention (Jain and Wallace, 2019), hierarchical document-level attention succeeded in identifying documents containing relevant evidence. The architecture here differs in that it builds representations hierarchically from the word level, as opposed to pre-extracted conceptual features, and takes document ordering into account using a bi-directional GRU (Bahdanau et al., 2015).

Specifically, our model has five layers (Figure 4). The first is a word-embedding layer that turns a one-hot word vector into a dense vector. The second to fourth layers are three Seq2Vec layers with attention that learn to aggregate, respectively, a sequence of word vectors into a sentence vector, a sequence of sentence vectors into a document vector, and a sequence of document vectors into an individual vector (hence 3HAN). The final layer is a fully connected layer followed by softmax.

We detail our Seq2Vec layer in the context of aggregating a sequence of document vectors to an individual's vector, though the three Seq2Vec layers are the same. See Figure 4b for an illustration. Document vectors $\{d_{i,j}\}_{j=1}^{m}$ are first passed through a bi-directional GRU layer. The outputs, after passing through a fully-connected layer and a non-linear layer, are then compared to a learnable attention vector, $v_{\text{attention}}$. Specifically,

$$g_{i,j} = \text{Bi-GRU}(d_{i,j}) \tag{8}$$

$$r_{i,j} = \tanh\left(W g_{i,j} + b\right) \tag{9}$$

$$a_{i,j} = \frac{e^{r_{i,j}^{\top} v_{\text{attention}}}}{\sum_{j'=1}^{m} e^{r_{i,j'}^{\top} v_{\text{attention}}}} \tag{10}$$

$$u_i = \sum_{j=1}^{m} a_{i,j} g_{i,j} \tag{11}$$

where $a_{i,j}$ is the normalized document attention score for the j-th vector, and u_i is the final aggregated individual vector. As shown in Equation 10, the transformed vector $r_{i,j}$ is compared with the learnable attention vector $v_{\text{attention}}$ using a dot product, and further normalized for the weighted averaging step in Equation 11.

Once we have the individual vector u_i, we can predict the risk label of the individual by passing it through a fully-connected layer and a softmax. Specifically,

$$P(\hat{y}_i) = softmax\left(W_{FC} u_i + b_{FC}\right) \tag{12}$$

Finally, we compare with the ground truth label y_i of individual i using negative log-likelihood to calculate a loss:

$$loss_i = -log\left(P\left(\hat{y}_i = y_i\right)\right). \tag{13}$$

5 Experimentation

We first introduce the test collection and then show how we can evaluate 3HAN and the cascade model baselines on the test collection using hTBG.

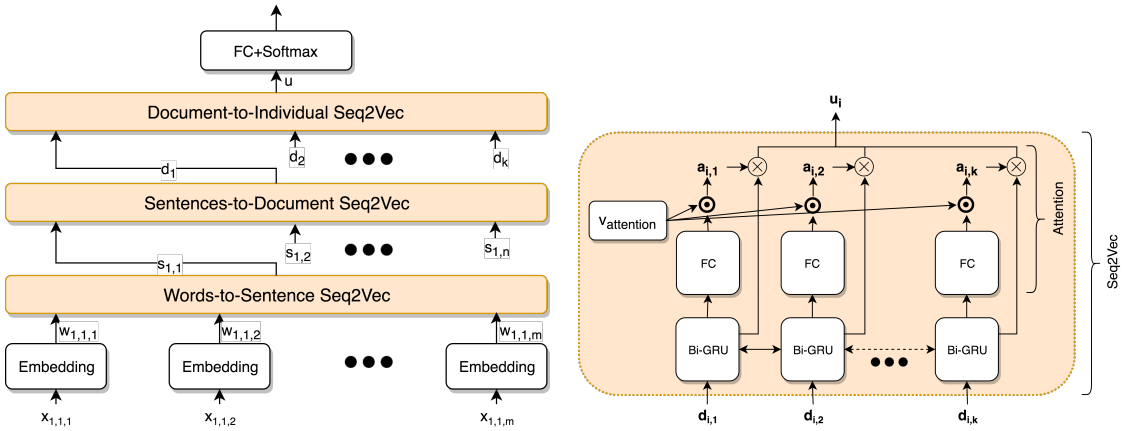
To demonstrate the effectiveness of the 3HAN model, which jointly learns to rank individuals and, within each individual, their posts as evidence, we compare it with different combinations of individual-level rankers and document-level rankers. Training details for all the models can be found in Appendix C.

5.1 Test Collection

In our experimentation, we use the University of Maryland Reddit Suicidality Dataset, v.2 (Shing et al., 2018; Zirikly et al., 2019).[6] This English-language dataset, derived from the 2015 Full Reddit Submission Corpus (2006-2015), includes 11,129 potentially at-risk individuals who posted on r/SuicideWatch (a subreddit dense in self-reports about suicidality, henceforth SW), as well as 11,129 control individuals who never posted on any mental-health related subreddit. Entire posting histories (not just from SW, but all Reddit forums) were collected.[7] An individual's number of posts can range from 10 to 1,326. See Table 2 for a detailed breakdown of number of posts per individual across datasets and risk categories.

The full dataset has three subsets with disjoint individuals. The first, which we term the WEAK SUPERVISION dataset, includes 10,263 individuals who posted in SW and 10,263 control individuals who did not; they are respectively considered to be indirectly positively and negatively labeled, very noisily since posting on SW does not necessary imply suicidal ideation.[8] The second set is the CROWDSOURCE dataset, including 621 individuals annotated by crowdsourcers with four risk levels: *No Risk*, *Low Risk*, *Moderate Risk*, and *Severe Risk*.

[6] See Appendix A for IRB and ethical considerations.
[7] See Gaffney and Matias (2018) for caveats.
[8] E.g. seeking help for a friend, or offering support.



(a) 3HAN (b) Seq2Vec with Attention

Figure 4: An illustration of the three-level Hierarchical Attention Network (3HAN) model

	# Posts	10-20	20-40	40-60	60-100	100-200	200-500	500-1,000	1,000-1,500
CrowdSource	No Risk	31	42	25	27	18	12	4	0
	Low Risk	19	22	5	11	2	4	0	0
	Moderate Risk	46	45	19	14	9	7	1	0
	Severe Risk	80	79	37	19	28	12	3	0
Expert	No Risk	3	7	2	5	7	8	3	0
	Low Risk	6	11	5	11	8	7	1	1
	Moderate Risk	23	19	12	26	13	14	5	3
	Severe Risk	7	2	5	9	10	4	4	1

Table 2: Number of individuals with the number (range) of posts, by dataset and risk category.

The last is the EXPERT dataset, including 242 individuals with the same four-level annotation, by four suicide risk assessment experts.[9] Along with the level of risk for each individual, the expert annotators also designated the single post that most strongly supported each of their low, moderate, or severe risk labels.

5.2 Evaluating with hTBG

As TBG and hTBG are measures designed for binary relevance judgements, we map the *Severe Risk* category to *at-risk*, and everything else to *not-at-risk*.[10] For word counts, we directly use the token counts in documents. We use the parameters that Smucker and Clarke (2012) estimated for TBG in user studies (Table 1). As discussed in Section 3.2, we assume there exists a maximum number of documents the clinician can read for each individual. We set that number to 50 for the calculation of hTBG; if no relevant document exists in the top 50 documents, we consider that individual a miss and set the gain to zero.[11]

To rank individuals using our classification models, we use a standard conversion method to convert four-class probability to a single score:

$$\sum_{\text{rel}_i}^{R} P\left(\hat{y}_i = \text{rel}_i\right) \text{score}_{\text{rel}_i} \qquad (14)$$

where R is $\{\text{No}, \text{Low}, \text{Moderate}, \text{Severe}\}$, and $\text{score}_{\text{rel}_i}$ is the real number that maps to the risk-level of the individual i. We use $\{\text{No} = 0, \text{Low} = 1, \text{Moderate} = 2, \text{Severe} = 4\}$ as our mapping — *No Risk* can plausibly be treated the same as a post with no annotation (e.g. a control individual), and exponential scaling also seems plausible although just one of many possibilities, which we leave for future work.

The hTBG metric also requires a stopping probability for each document, $R_{i,l}$. Assuming that the more severe the risk associated with a document is, the more likely the assessor is to stop and flag the

[9] Shing et al. (2018) report reliable expert annotation, Krippendorff's $\alpha = .81$. The original EXPERT dataset had 245 individuals; we exclude three owing to errors in processing.

[10] Since the label definitions distinguish severe from moderate by focusing on the risk of an attempt *in the near future*, this binary distinction is aligned with recent work in suicidology that focuses specifically on characterizing "the acute mental state that is associated with near-term suicidal behavior" (Schuck et al., 2019).

[11] All parameters were frozen prior to testing. We plan to estimate hyperparameters in our own user studies in the future.

individual, on the EXPERT dataset where we have document-level annotations, we can estimate the expected stopping probability as:

$$R_{i,l} = 1 - \prod_{c=1}^{C} \left(1 - \frac{\text{score}_{\text{rel}_{i,l,c}}}{\text{score}_{\text{max}}}\right) \quad (15)$$

where C annotators annotated the post as most strongly supporting their judgment. $\text{Score}_{\text{rel}_{i,l,c}}$ is a mapping from the document-level risk by annotator c to a real number, with the same mapping used in Equation 14. $\text{Score}_{\text{max}} = 4$ is the maximum in that mapping.

To reflect different time budgets, we report results with the half-life parameter ranging from 1 to 6 hours, which corresponds to expected reading time budgets from 1.4 to 8.7 hours.

5.3 Models for Ranking Individuals

3HAN. 3HAN is first pretrained on the binary WEAK SUPERVISION dataset. The model is then further tuned on the four-class CROWDSOURCE dataset by transferring the weights (except the last fully-connected prediction layer) over. We initialized and fixed the word embedding using the 200-dimensional Glove embedding trained on Twitter (Pennington et al., 2014).[12]

3HAN_Av. 3HAN Average is trained the same way as 3HAN, except that the last Seq2Vec layer (the layer that aggregates a sequence of document vectors to an individual vector) is averaged instead of using attention, which can be achieved by fixing $a_{i,j} = \frac{1}{m}$ in Equation 10. This is similar to the HN-AVE baseline in Yang et al. (2016). Note that 3HAN AV cannot rank documents, as it lacks document attention.

LR. A logistic regression model is trained on the CROWDSOURCE dataset. The feature vector for an individual is computed by converting documents into document-level feature vectors, and then averaging them to obtain an individual-level feature vector. For each document, we concatenate four feature sets: (1) bag-of-words for vocabulary count larger than three, (2) Glove embedding summing over words, (3) 194 features representing emotional topics from Empath (Fast et al., 2016), and (4) seven scores measuring document readability.[13] This model is included as a conventional baseline in suicide risk assessment, similar to the baseline found in Shing et al. (2018).

5.4 Models for Ranking Documents

3HAN_Att. Document attention learned jointly with 3HAN. As a side effect to training our 3HAN model, we learn document attention scores, see Equation 10. This score can then be used to rank documents in terms of their relevance to the judgement. This availability of document ranking, despite a lack of document annotations, is a significant advantage of hierarchical attention networks, since fine-grained document annotations are difficult to obtain on a large scale. Sentence- and word-level attention are a further advantage, in terms of potentially facilitating user review (see Figure 1), although exploring that awaits future work.

Forward and Backward. Ranking an individual's documents in either chronological order or reverse chronological order is an obvious default in the absence of a trained model for document ranking, important baselines for testing whether a document ranking model actually adds value.

6 Results and Discussion

Our model, 3HAN+3HAN_ATT, the only joint model, achieves the best performance on hTBG compared to all other combinations of individual rankers and document rankers across three different time budgets (Table 3). The result is significant except when compared to 3HAN_AV+3HAN_ATT.[14] However, using 3HAN_ATT to rank documents implies that you have already trained 3HAN. Therefore, a more reasonable combination to compare with is 3HAN_AV+BACKWARD, which we outperform by a significant margin.

Overall, the effect of document ranking is larger than the effect of individual ranking. Notably, the FORWARD document ranker always yields the worst performance. BACKWARD, on the other hand, is surprisingly competitive. We hypothesize that this may be an indication that suicidal ideation worsens over time, or perhaps of the unfortunate

[12] We experimented with trainable Glove embedding as well as BERT, but saw little to no improvement in performance using cross-validation. We plan to explore fine-tuning BERT on Reddit in future work.

[13] Flesch-Kincaid Grade Level, Flesch Reading Ease, Dale Chall Readability, Automated Readability Index (ARI), Coleman Liau Index, Gunning Fog Index, and Linsear Write.

[14] Paired bootstrap resampling test, repeated 1000 times, $p < 0.05$.

Individual Ranker	Document Ranker	Half-life h 1 hr	3 hrs	6 hrs
LR	FORWARD	7.51	10.05	10.89
3HAN_AV	FORWARD	7.76	10.15	10.94
3HAN	FORWARD	7.40	9.98	10.84
LR	BACKWARD	8.75	11.70	12.68
3HAN_AV	BACKWARD	9.65	12.09	12.89
3HAN	BACKWARD	9.73	12.17	12.95
LR	3HAN_ATT	9.44	12.05	12.88
3HAN_AV	3HAN_ATT	10.16	12.35	13.04
3HAN	3HAN_ATT	**10.39**	**12.49**	**13.12**
Optimal hTBG		19.78	20.39	20.54

Table 3: hTBG scores with three different time budgets, all combinations of individual and document rankers.

event of suicide attempts following posting a *Severe Risk* document. This motivates the importance of prioritizing the reading order of documents: being able to find evidence early in suicide assessment leaves more time for other individuals, and will reduce probability of misses.

Document ranking alone does not decide everything, as 3HAN+BACKWARD outperforms LR+3HAN_ATT. It is the combination of 3HAN and its document attentions that produce our best model. This makes sense, as 3HAN, while learning to predict the level of risk, also learns which documents are important to make the prediction.

Figure 1 shows the top 3 documents in a summary-style view for each of the highest ranked 3 individuals, with word-level attention shown using shading. Words without attention are obfuscated; others are altered to preserve privacy.

Previously Existing Measures. For previously existing measures, e.g. TBG and NDCG@20, document ranking has no effect, and thus these are not suitable measures in our scenario. However, we include results here for reference (Table 4). Since 3HAN AV. and LR cannot rank documents, it is impossible to calculate hTBG, so we report results on the chronologically backward ranking strategy. NDCG@20 is NDCG score cut off at 20, chosen based on the optimal hTBG value.

7 Conclusions and Future Work

We introduced hTBG, a new evaluation measure, as a step toward moving beyond risk classification to a paradigm in which prioritization is the focus, and where time matters. Like TBG, the hTBG score is interpretable as a lower bound on the expected number of relevant items found in a ranking, given a time budget. In our experiment, a "relevant item" is a person classified by experts as being at risk of attempting suicide in the near future.

Ranker	hTBG	TBG	NDCG@20
3HAN+3HAN ATT.	**12.49**	11.46	70.90
3HAN AV.+BACKWARD	12.09	11.40	68.28
LR+BACKWARD	11.70	10.98	69.44
Optimal	20.39	19.75	100.00

Table 4: TBG and NDCG@20 listed to compare with hTBG. Both hTBG's and TBG's half lives are set at 3 hrs, and maximum document cutoff is set at 50.

Measured at an expected reading time budget of about half a day (4hr20min, half-life 3hrs), our joint ranking approach achieved hTBG of 12.49 compared with 11.70 for a plausible baseline from prior art: using logistic regression to rank individuals, and then looking at a individual's posts in backward chronological order. That increase is just a bit short of identifying one more person in need of immediate help in the experiment's population of 242 individuals. There are certainly limitations in our study and miles to go before validating our approach in the real world, but our framework should make it easy to integrate and explore other individual rankers, document rankers and explanation mechanisms, and to actually build user interfaces like the schematic in Figure 1.

Acknowledgments

This work has been supported in part by a University of Maryland Strategic Partnership (MPower) seed grant, an AWS Machine Learning Research Award, and an AI + Medicine for High Impact (AIM-HI) Challenge Award. We are immensely grateful to Glen Coppersmith, Michelle Colder Carras, April Foreman, Michelle Kuchuk, Beau Pinkham, Rebecca Resnik, Katherine Musacchio Schafer, Jonathan Singer, Raymond Tucker, Tony Wood, Ayah Zirikly, members of the UMIACS CLIP lab, and participants at the Workshops on Computational Linguistics and Clinical Psychology for valuable discussions related to this work.

References

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*,

WSDM '09, page 5–14, New York, NY, USA. Association for Computing Machinery.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Krisztian Balog. 2018. *Entity-Oriented Search*, volume 39 of *The Information Retrieval Series*. Springer.

Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov, and Luo Si. 2012. Expertise retrieval. *Foundations and Trends in Information Retrieval*, 6(2–3):127–256.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2018. Multi-label classification of patient notes: Case study on ICD code assignment. In *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence*, pages 409–416. AAAI Press.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL*, pages 94–102. Association for Computational Linguistics.

David E. Bloom, Elizabeth Cafiero, Eva Jané-Llopis, Shafika Abrahams-Gessel, Lakshmi Reddy Bloom, Sana Fathima, Andrea B. Feigl, Tom Gaziano, Ali Hamandi, Mona Mowafi, Danny O'Farrell, and Emre. 2012. The Global Economic Burden of Noncommunicable Diseases. PGDA Working Papers 8712, Program on the Global Demography of Aging.

Bureau of Health Workforce. 2020. Designated health professional shortage areas: Statistics, second quarter of fiscal year 2020, designated HPSA quarterly summary.

Rafael A. Calvo, David N. Milne, M. Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Nat. Lang. Eng.*, 23(5):649–685.

Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009*, pages 621–630. ACM.

Munmun De Choudhury. 2013. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd international workshop on Socially-aware multimedia, SAM@ACM Multimedia 2013*, pages 49–52. ACM.

Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication*, 1:343–359.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, 10:117822261879286.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM.

Joseph C. Franklin, Jessica D. Ribeiro, Kathryn R. Fox, Kate H. Bentley, Evan M. Kleiman, Xieyining Huang, Katherine M. Musacchio, Adam C. Jaroszewski, Bernard P. Chang, and Matthew K. Nock. 2017. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, 143(2):187–232.

Devin Gaffney and J. Nathan Matias. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *PLOS ONE*, 13(7):1–13.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6. Association for Computational Linguistics.

Holly Hedegaard, Sally C Curtin, and Margaret Warner. 2018. Suicide rates in the United States continue to increase. *National Center for Health Statistics*.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 1373–1378. The Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 3543–3556. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Kathryn P. Linthicum, Katherine Musacchio Schafer, and Jessica D. Ribeiro. 2019. Machine learning in suicide science: Applications and ethics. *Behavioral Sciences & the Law*, 37(3):214–222.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. eRisk 2020: Self-harm and depression challenges. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 557–563. Springer.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2016*, pages 118–127. The Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, pages 1101–1111. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543. ACL.

Tetsuya Sakai. 2019. Graded relevance assessments and graded relevance measures of NTCIR: A survey of the first twenty years. *CoRR*, abs/1903.11272.

SAMHSA. 2019. *National Survey on Drug Use and Health, 2017 and 2018*. Center for Behavioral Health Statistics and Quality. Table 8.58B.

Allison Schuck, Raffaella Calati, Shira Barzilay, Sarah Bloch-Elkouby, and Igor Galynker. 2019. Suicide Crisis Syndrome: A review of supporting evidence for a new suicide-specific diagnosis. *Behavioral sciences & the law*, 37(3):223–239.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, CLPsych@NAACL-HTL*, pages 25–36. Association for Computational Linguistics.

Han-Chin Shing, Guoli Wang, and Philip Resnik. 2019. Assigning medical codes at the encounter level by paying attention to documents. In *ML4H, Machine Learning for Health Workshop at NeurIPS*.

Mark D. Smucker and Charles L. A. Clarke. 2012. Time-based calibration of effectiveness measures. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12*, pages 95–104. ACM.

Ellen M. Voorhees. 2001. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer.

Byron C. Wallace. 2019. Thoughts on "attention is not not explanation". Medium, Accessed: December, 2019.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 11–20. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. The Association for Computational Linguistics.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978.

Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, 12(4):313–325.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33. Association for Computational Linguistics.

A Appendix: Ethical Considerations

Our research involving the University of Maryland Reddit Suicide Dataset has undergone review by the University of Maryland Institutional Review Board with a determination of Category 4 Exempt status under U.S. federal regulations. For this dataset, (a) the original data are publicly available, and (b) the originating site (Reddit) is intended for anonymous posting. In addition, since Reddit is officially anonymous, but that is not enforced on the site, the dataset has undergone automatic de-identification using named entity recognition aggressively to identify and mask out potential personally identifiable information such as personal names and organizations, in order to create an additional layer of protection (Zirikly et al., 2019). In an assessment of de-identification quality, we manually reviewed a sample of 200 randomly selected posts (100 from the SuicideWatch subreddit and 100 from other subreddits), revealing zero instances of personally identifiable information.

Following Benton et al. (2017), we treat the data (even though de-identified) as sensitive and restrict access to it, we use obfuscated and minimal examples in papers and presentations, and we do not engage in linkage with other datasets.

The dataset is available to other researchers via an application process put in place with the American Association of Suicidology that requires IRB or equivalent ethical review, a commitment to appropriate data management, and, since ethical research practice is not just a matter of publicly available data or even IRB approval (Zimmer, 2010; Benton et al., 2017; Chancellor et al., 2019), a commitment to following additional ethical guidelines. Interested researchers can find information at http://umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html.

B Appendix: Proofs

B.1 Time-Biased Gain

In order to prove that TBG statisfies the *speed-biased* criterion, consider two individuals ranked at consecutive positions k and $k+1$; if we swap the two individual, the change in TBG score is:

$$\begin{aligned}\Delta\text{TBG} = &(g_{k+1} - g_k)D(T(k)) \\ &+ g_k D\left(T(k) + t(k+1)\right) \\ &- g_{k+1} D\left(T(k) + t(k)\right)\end{aligned} \tag{16}$$

This leads to Lemma B.1-B.3:

Lemma B.1. *Swapping a not-at-risk individual ranked at k with an at-risk individual ranked at $k+1$ always increases TBG.*

Proof. Let $g_k = 0$ and $g_{k+1} > 0$. Equation 16 simplifies to

$$\Delta\text{TBG} = g_{k+1}\left(D(T(k)) - D(T(k) + t(k))\right) \tag{17}$$

which is always positive because the decay function monotonically decreases, and each assessment of an individual requires at least T_s seconds. □

Lemma B.2 (Risk-based Criterion)**.** *The optimal value of TBG under binary relevance is obtained only if all not-at-risk individuals are ranked below all at-risk individuals.*

Proof. Let π be a ranking of individuals that yields the optimal value of TBG. Assume that in π there exist not-at-risk individuals ranked before at-risk individuals. Let the k-position be the lowest ranked not-at-risk individual that is at least in front of one at-risk individual, we can then apply Lemma B.1 to increase TBG. This leads to a contradiction. □

Lemma B.3. *Swapping an at-risk individual of longer assessment time ranked at k of with an at-risk individual of shorter assessment time ranked at $k+n$, where $k+n$ is the closest at-risk individual ranked lower than k, always increases TBG.*

Proof. Let $g_k = g_{k+n} > 0$, and $\forall i \in \{i | k < i < k+n\}, g_i = 0$. We have

$$\begin{aligned}\Delta\text{TBG} = g_k(&D(T(k+n) + t(k+n) - t(k)) \\ &- D(T(k+n)))\end{aligned} \tag{18}$$

which is always positive because the decay function monotonically decreases, and $t(k+n) < t(k)$ from the assumption that the individual at $k+n$ has shorter assessment time. □

Lemma B.3 naturally leads to a proof for the speed-biased property of TBG:

Proof for Theorem 3.1. Applying Lemma B.3, we know that swapping k and $k+r$ leads to a positive gain between the two. Now, consider all

at-risk individuals ranked between k and $k+r$: $\forall u$, s.t. $k < u < k+r$, the difference is:

$$g_u(D(T(u)+t(k+r)-t(k))-D(T(u))) \quad (19)$$

which is always greater than or equal to zero due to the fact that the decay function monotonically decrease, and $t(k+r) < t(k)$. Thus, the net difference is always larger than zero, thus satisfying the *speed-biased* criterion. □

Finally, combing previous results, we can easily show:

Proof for Theorem 3.2. A direct consequence of Theorem 3.1 is that if the at-risk individuals are sorted by assessment time in ascending order, no swapping between any two individuals can increase TBG. This, combined with Lemma B.2, that all at-risk individuals are on top of not-at-risk individuals, leads to the necessary condition. Because any swapping within the not-at-risk individuals does not change TBG when no at-risk individuals are ranked lower, this implies that ranking according to Theorem 3.2 gives us a unique and optimal value, which satisfies the sufficient condition of Theorem 3.2. □

B.2 Hierarchical Time-Biased Gain

The assessment time of an individual ranked at k, $t(k)$, is monotonic with E_i, thus showing minimal value of E_i suffices. Recall that E_i is calculated as:

$$E_i = T_\alpha \sum_{l=1}^{L} \left(W_{i,l} \prod_{m=1}^{l-1} (1 - R_{i,m}) \right) + T_\beta \quad (20)$$

Consider, again, swapping a document at rank l with a document at rank $l+1$ belonging to the same individual i. The change in E_i is:

$$\Delta E_i = \kappa_{i,l} \left(W_{i,l+1} R_{i,l} - W_{i,l} R_{i,l+1} \right) \quad (21)$$

where $\kappa_{i,l} = T_\alpha \prod_{j=1}^{l-1} (1 - R_{i,j}) \geq 0$ is a fixed term that is not affected by the swap.

Equation 21 also points to an important observation:

Lemma B.4. *If $W_{i,l+1} R_{i,l} - W_{i,l} R_{i,l+1} < 0$ and $R_{i,j} < 1$ for all $j < l$, then swapping document l with document $l+1$ will decrease E_i.*

Proof. This follows directly from Equation 21. □

Lemma B.5. *If $R_{i,j} < 1$ for all j, then minimum individual assessment time is obtained if and only if the documents are sorted in descending order by*

$$\frac{R_{i,l}}{W_{i,l}}. \quad (22)$$

Proof. Let τ be a document ranking that yields the minimum individual assessment time, and for the sake of contradiction, not a ranking that can be obtained by ranking according to $\frac{R_{i,l}}{W_{i,l}}$. We can, thus, find two neighboring documents, without loss of generality, l and $l+1$, such that:

$$\frac{R_{i,l}}{W_{i,l}} < \frac{R_{i,l+1}}{W_{i,l+1}} \quad (23)$$

this leads to:

$$R_{i,l} W_{i,l+1} - R_{i,l+1} W_{i,l} < 0 \quad (24)$$

since all $W > 0$. Lemma B.4 together with the prerequisite that $R_{i,j} < 1$ for all j then suggest that swapping the two leads to a decrease of E_i. This contradicts with the assumption that τ is an optimal ranking. This proves that to achieve minimum individual assessment time, it is necessary to sort by $\frac{R_{i,l}}{W_{i,l}}$. The sufficient condition follows by the fact that swapping tied documents does not lead to change in E_i, as shown in Equation 21 □

Proof for Theorem 3.3. Let τ be a document ranking according to $\frac{R_{i,l}}{W_{i,l}}$. Let m be the document such that $R_{i,m} = 1$ and is ranked closer to the top then any other document with $R_{i,:} = 1$ (i.e. with the shortest $W_{i,:}$). Now, consider using m to cut the documents into two partitions: the first partition of documents are ones ranked before m. Applying Lemma B.5, this partition of documents are already in optimal sorted order, since there's no $R_{i,:} = 1$. The second partition, documents ranked lower than m, the ranking simply does not matter, as Equation 20 shows, the $(1 - R_{i,m})$ term will make everything zero afterwards.

Now, let's consider moving a document from the second partition to the first partition. Since any documents in the second partition has a $\frac{R_{i,j}}{W_{i,j}}$ that is smaller than any documents in the first partition, after you move the document, the optimal ranking for the first partition will put the document at the bottom, right next to m. And since $\frac{R_{i,m}}{W_{i,m}} \geq \frac{R_{i,j}}{W_{i,j}}$ due to the original ordering, we can apply Lemma B.4, which can swap the document back below m. Next,

consider moving the lowest ranked document of the first partition (the one ranked at $m-1$) to the second partition. This will always increase E_i, as shown from Lemma B.4. Moving any other document in the first partition will also increase E_i as least as much as before, since the process is equivalent to swapping with (and thus potentially increase E_i) any intermediate documents in between.

Combine these two together, we show that E_i is at a minimum value when sorted in descending order according to $\frac{R_{i,l}}{W_{i,l}}$. □

B.3 Relationship between ERR and hTBG

Here we show the derivation from the cascading user model in ERR to the individual assessment time estimation (E_i) in hTBG. ERR assumes a stopping probability (written in hTBG terms):

$$P(\text{stop at l}) = R_{i,l} \prod_{j=1}^{l-1} (1 - R_{i,j}) \tag{25}$$

The expected words read, can then be calculated as:

$$\begin{aligned} &\sum_{l=1}^{L} \left(P(\text{stop at } l) \sum_{d=1}^{l} W_{i,l} \right) \\ =&\sum_{l=1}^{L} \left(R_{i,l} \prod_{j=1}^{l-1} (1 - R_{i,j}) \left(\sum_{d=1}^{l} W_{i,l} \right) \right) \end{aligned} \tag{26}$$

This can be rearranged to the formula we used in hTBG:

$$\sum_{l=1}^{L} \left(W_{i,l} \prod_{m=1}^{l-1} (1 - R_{i,m}) \right) \tag{27}$$

by letting $R_{i,L} = 1$ (the user has to stop reading at the last document). To show this, observe that $W_{i,1}$ appears in all L terms of the summation, thus the coefficient for $W_{i,1}$ is simply $\sum_{l=1}^{L} (R_{i,l} \prod_{j=1}^{l-1} (1 - R_{i,j})) = 1$, from both simple manipulation and the fact that we are summing over probability. Similarly, $W_{i,2}$ appears in all L terms except with $l = 1$, thus $(1 - R_{i,1})$. For $W_{i,3}$ it is $(1 - R_{i,1}) - R_{i,2}(1 - R_{i,1}) = \prod_{j=1}^{2} (1 - R_{i,j})$. The rest follows.

C Appendix: Training Details

All models are built using AllenNLP (Gardner et al., 2018). Tokenization and sentence splitting are done using spaCy (Honnibal and Johnson, 2015).

The CROWDSOURCE dataset is split into a training set (80%) and a validation set (20%) during model development. We did not test on the EXPERT dataset until all parameters of the models were fixed. Cross validation on the training set is used for hyperparameter tuning. For 3HAN, we used ADAM with learning rate 0.003, trained for 100 epochs with early stopping on the validation dataset, with patience set to 30. For 3HAN_AV, the same hyperparameters are used. For LR, we used SGD with learning rate 0.003, trained for 100 epochs with early stopping on the validation dataset, with patience set to 30.

Both 3HAN and 3HAN_AV's Seq2Vec layers use bi-directional GRU with attention. The word-to-sentence layer has input dimension of 200, hidden dimension of 50, and output dimension of 100, since the GRU is bi-directional. The sentence-to-document and document-to-individual layer, similarly, has input dimension of 100, hidden dimension of 50, and output dimension of 100. Hyperparameters were selected using cross validation on the training set split of the CROWDSOURCE dataset.