To Comprehend the *New*: On Measuring the Freshness of a Document

Tirthankar Ghosal
IIT Patna
Patna, India
tirthankar.pcs16@iitp.ac.in

Abhishek Shukla
IIIT Kalyani
Kalyani, India
abhishek@iiitkalyani.ac.in

Asif Ekbal
IIT Patna
Patna, India
asif@iitp.ac.in

Pushpak Bhattacharyya
IIT Patna
Patna, India
pb@iitp.ac.in

***Abstract*—Detecting the novelty or freshness of an entire document is essential in this age of data duplication and semantic-level redundancy all across the web. Current techniques for the problem mostly root on handcrafted similarity and divergence based measures to classify a document as novel or non-novel. However, document-level novelty detection is relatively less explored in literature if compared to its sentence-level counterpart. In this work, we present a deep neural architecture to automatically predict the amount of new information contained in a document in the form of a novelty score. Along with, we offer a dataset of more than 7500 documents, annotated at the sentence-level to facilitate further research. Our approach which learns the notion of novelty and redundancy only from the data achieves significant performance improvement over the existing methods and adopted baselines ($\sim$ 17% error reduction). Also, our approach complies with the Two-Stage theory of human recall essential to comprehend new information.**

***Index Terms*—novelty score, document-level novelty, document classification**

I. Introduction

Novelty detection from texts implies identifying or retrieving relevant pieces of texts that carry new information and has not been previously seen or known to the reader. The document-level variant of the problem is a task of categorising a document (as *novel, non-novel or partially novel*) based on the amount of new information contained in the document[1]. Although sentence-level novelty detection is a well-diagnosed problem in information retrieval literature, we find a very little amount of work on novelty detection at the document level (See Section II). Moreover, the research on the concerned problem encompassing semantic-level comprehension of documents is pretty hard to find. Maybe because every document contains something in new [1]. Comprehending the novelty of an entire document with confidence is even a very complex task for humans. Robust semantic representation of documents is still an active area of research which also somewhat limits the investigation of novelty mining at the document-level. Categorising a document as novel or non-novel is not that straightforward and involves complex semantic phenomena of inference, relevance, diversity, relativity, temporality as is shown in [2]. But considering the exponential rise of documents all across the web, automatically figuring out the semantically redundant document(s) seems an important problem to probe.

[1]w.r.t a set of relevant documents already seen by the reader

According to a report from Google[2] and a certain SEO study[3], about 25-30% of the web's content is duplicate. Hence filtering out superfluous documents and identifying original ones are essential. More important is to see how much new information a particular document carries. The appetite of novel information is different with different readers [3] which makes the problem a very subjective one. Hence figuring out the amount of new information in a document and then letting the users decide the category of the document (novel, non-novel, partially novel) appears an excellent direction to investigate into this problem. In our current work, we study this seemingly interesting problem: *to predict the novelty score of a document based on the document(s) or information already seen by the system*. We develop an appropriate dataset for the problem and apply our methods to reach as close as possible to the ground truth. We design a deep neural architecture leveraging on the power of natural language inference to understand the notion of new and redundant information, only from the data (without explicitly specifying features/rules). Our results show that we outperform the standard baselines and *state-of-the-art* by a wide margin in terms of various metrics (as shown in Table III).

II. Related Work

Novelty detection from texts came into light with the First Story Detection (FSD) task in the Topic Detection, and Tracking (TDT) exercises [4]. Some notable approaches with the TDT benchmark were by [5]–[8]. It was first here that redundancy (similarity) came out as an opposite characteristic to novelty. It became popular to consider documents having similarity score less than a certain threshold as novel. The problem gained impetus with inclusion in the tracks of TREC [9] evaluation exercises from 2002-2004. Although, the focus was on novel sentence retrievals, several methods came up including that of [10]–[14]. The novelty detection subtask in the RTE-TAC 2010, 2011 [15] is the first to correlate textual entailment as one approximation to non-novelty.

At the document level, [16] first studied the role of information filtering systems to identify relevant and novel documents and used measures like set difference, cosine similarity, distributional similarity with Dirichlet and shrinkage smoothing

[2]https://www.youtube.com/watch?v=mQZY7EmjbMA

[3]https://raventools.com/studies/onpageseo/#duplicate

and a mixture model on their APWSJ corpus. Reference [17] investigated asymmetric overlap measures on the TREC 2004 novelty detection dataset and APWSJ. Reference [18] explored novelty scoring, where they convert similarity score into a novelty score by setting: *Novelty Score= 1-Similarity Score*. Reference [19] proposed blended metrics for novelty scoring combining both cosine similarity and new word ratio. Reference [20] explored novelty scoring using language modelling via KL Divergence. Reference [21] used Inverse Document Frequency as one approximation to deduce the novelty score of a document with respect to a corpus. Quite recently [22] formulated an entropy-based model to score the innovativeness of textual ideas automatically.

However, our approach is fundamentally different from all these approaches in the sense that, we do not carve any rule or feature to estimate the novelty score. Instead, we let our deep neural architecture understand the notion of new information and redundancy only from the data.

III. Problem Definition

We define the problem as associating a qualitative novelty score to a document based on the amount of new information contained in it. Let us consider the following example:

Source Text: *Singapore, an island city-state off southern Malaysia, is a global financial center with a tropical climate and multicultural population. Its colonial core centers on the Padang, a cricket field since the 1830s and now flanked by grand buildings such as City Hall, with its 18 Corinthian columns. In Singapore's circa-1820 Chinatown stands the red-and-gold Buddha Tooth Relic Temple, said to house one of Buddha's teeth.*

Target Text: *Singapore is a city-state in Southeast Asia. Founded as a British trading colony in 1819, since independence it has become one of the world's most prosperous, tax-friendly countries and boasts the world's busiest port. With a population size of over 5.5 million people it is a very crowded city, second only to Monaco as the world's most densely populated country.*

The task is to find the novelty score of the target text w.r.t the source text. It is quite clear that the target text is having new information with respect to the source except that the first sentence in the target contains some redundant content (*Singapore is a city-state*). Analysing the first sentence in the target text we get two information: that *Singapore is a city-state* and *Singapore lies in Southeast Asia*. Keeping the source text in mind, we understand that the first part is *redundant* whereas the second part has new information, i.e., we can infer that 50% information is novel in the first target sentence. Here, we consider only the surface-level information in the text and do not take into account any pragmatic knowledge of the reader regarding the geographical location of Singapore and Malaysia in Asia. Here, our new information appetite is more fine-grained and objective.

Now let us attach a qualitative score to each of the three target sentences as 0.5, 1.0, 1.0, signifying 50% new information (0.5) and total new information (1.0), respectively. The cumulative sum comes to 2.5 which says that the target text has 83.33% new information w.r.t the source text[4]. This scoring mechanism, although straightforward, intuitively resembles the human-level perception of the amount of new information. However, we do agree that this approach attaches equal weights to long and short sentences. Long sentences would naturally contain more information whereas short sentences would convey less information. Also, we do not consider the relative importance of sentences within the documents. However, for the sake of initial investigation and ease of annotation[5], we proceed with this simple quantitative view of novelty and create a dataset that would be a suitable testbed for our experiments to predict the document-level novelty score.

IV. Dataset Creation

As discussed earlier, [2] developed a new dataset for document-level novelty detection (TAP-DLND 1.0). Keeping the dataset structure identical (Figure 1), we extend the dataset and re-annotate the extended dataset from scratch to incorporate our sentence to document-level novelty scoring perspective. The sentence-level annotation guidelines are entirely different from the document-level annotations in [2] (See Section IV-C, Figure 2). The current extended dataset statistics are in Table II.

A. TAP-DLND 1.0 Dataset Description

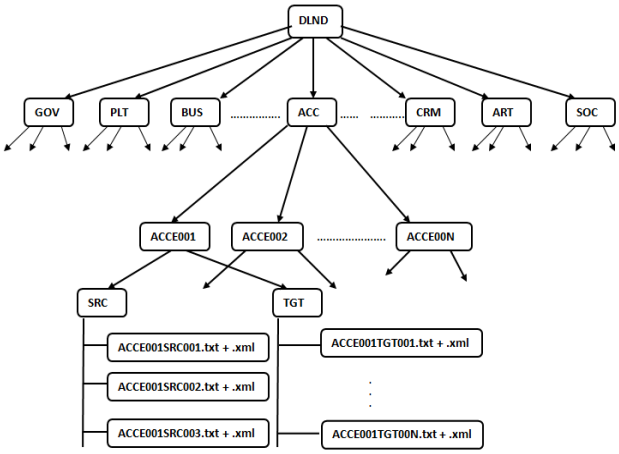

Fig. 1: The TAP-DLND 1.0 corpus structure. We retain the structure in the extended dataset we use in the current work

The TAP-DLND 1.0 tracks the event-level development of a news item and contains 10 different categories of news: Government (GOV), Crime (CRM), Arts and Entertainment

[4]if all the sentences were tagged as novel, the score would have been 3.0 indicating 100% novel information in the target text

[5]identifying and annotating an information unit would have been complex. However, we plan for further research with annotation at the phrase-level and with relative importance scores

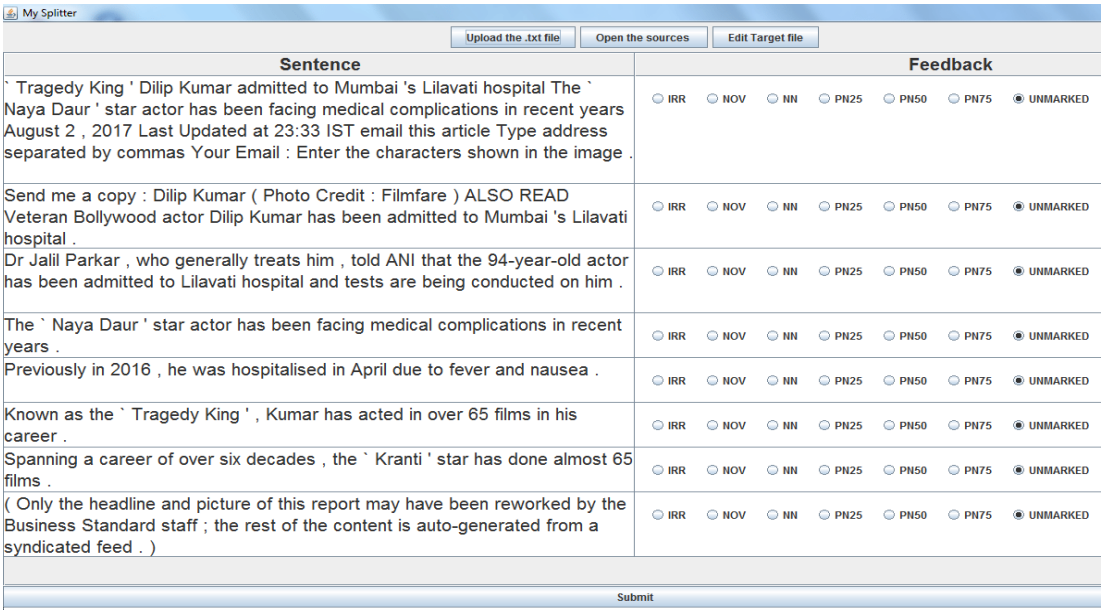

Fig. 2: The Sentence-Level Annotation Interface used to generate the Document-Level Novelty Score (Gold Standard)

(ART), Sports (SPT), Accident (ACC), Politics (PLT), Business (BUS), Nature (NAT), Society (SOC) and Terrorism (TER). Each category consists of several events and each event, in turn, consists of several documents describing the event in a chronological sequence denoting its development over time. The event-descriptor documents are crawled from various sources on the web in a temporal fashion. Now for each event, three documents (news reporting) were purposely chosen as the *source* and the rest as the *target*.

The source documents were intuitively selected from the initial dates of occurrence/reporting of a particular event. Care has been taken to choose the source documents such that they represent different facets of information regarding the same event, i.e., they cover mutually exclusive information. For a particular event, the source reportings are the seed documents which represents the memory or information already known/seen by the reader. The task is to determine the state of novelty or the quantity of new information in a target document against these designated source documents. The annotators were asked to read the source documents first, and then provide binary judgments (*novel,non-novel*) for each of the target documents. The intuition is: for a particular event in a given date (suppose the event of an accident), different news sources would report almost the same information and hence would be semantically non-novel to one another. However, updates on that event in subsequent dates would lead to novel information.

We re-annotate this dataset from scratch, but at the sentence level (also extend to more than 7500 documents), to deduce a document-level novelty score for each target document.

B. Why sentence level annotation?

The judgment of novelty at the discourse level is difficult to comprehend and is too much dependent on the understanding by the human subject. It is quite likely that every document may contain something new w.r.t. previously seen information [1]. However, this relative amount of new information is not always justified to deem the entire document as novel[6]. Hence, we deem that instead of looking at the target document in entirety, if we look into the sentential information content, we may get a more fine-grained objective view of new information content in a document discourse. Thus with this motivation, we formulate a new set of annotation guidelines to be followed while annotating at the sentence-level (in context to the discussion in Section III). We associate scores with each annotation judgment which finally cumulates to a document-level novelty score.

C. Annotation Schema

We design an easy to navigate interface (Figure 2) to facilitate the annotations and perform the annotation event-wise. For a particular event, an annotator reads the pre-determined three seed source documents, gathers information regarding that particular event and then proceeds to annotate the target documents[7], one at a time. Upon selection of the desired target document, the interface splits the document into constituent sentences[8] and allows six different annotation options for each target sentence (Table I). We finally take the cumulative average as the document-level novelty score for the target document. We exclude the sentences marked as irrelevant (IRR) from the calculation.

V. Methodology

Our main intention is to predict the novelty score of a document given a set of relevant documents already seen by

[6]it is where the sentence significance in the discourse comes to play which lies in the scope of our further research

[7]for one particular event there are three source documents but multiple target documents

[8]using NLTK sentence splitter

Annotation Labels	Description	Score
Novel (NOV)	The entire sentence has new information.	1.00
Non-Novel (NN)	The information contained in the sentence is redundant.	0.00
Little bit Novel (PN25)	The sentence has a little bit of new information. Most of the information is overlapping with the source.	0.25
Partially Novel/Non-Novel (PN50)	The sentence has an almost equivalent amount of new and redundant information	0.50
Mostly Novel (PN75)	Most of the information in the sentence is new	0.75
Irrelevant (IRR)	The sentence is irrelevant to the event/topic in context	—

TABLE I: Sentence-level annotations. These are w.r.t. the information contained in the source documents for each event. The annotations are qualitatively defined. We assign scores to quantify them (see the discussion in Section III).

Dataset Characteristics	Statistics
Event categories	10
Number of events	245
Number of source documents per event	3
Total target documents	7536
Total sentences annotated	**120,116**
Average number of sentences per document	$\sim$ 16
Average number of words per document	$\sim$ 385
Inter-rater agreement	0.88

TABLE II: Extended TAP-DLND 1.0 dataset statistics. Inter-rater agreement [23] is measured for 100 documents for sentence-level annotations by two raters.

the model. Having created an appropriate dataset, we design an architecture, which encodes the target document information jointly with the source information in one single unit and then makes use of a deep Convolutional Neural Network (CNN) to predict the novelty score[9]. This jointly encoded target document representation is particularly important here, because, for novelty, the context of the topic in concern plays a pivotal role.

A. Premise Selection

Selecting (Recalling) the appropriate source document(s) is essential for novelty search [24]. We relate this with the *Two-Stage theory* of human recall [25] consisting of **Phase-I: Search and Retrieval** and **Phase-II: Recognition**. Here to realise Phase-I: from the pool of all source documents (simulating the memory or information already known), we select the top 10 documents which could be the potential source of a given target document. We take Named-Entities similarity [26], [27] to retrieve the 10 most similar documents. The Recall@10 here at this stage is 0.93. For Phase-II, out of the retrieved 10, we further filter three potential source documents[10] via the Word Mover's Distance [28] [11]. Less is the distance; higher is the ranking for relevance. The Recall@3 at this stage is 0.94.

B. Source Encapsulated Target Document Vector (SETDV)

As discussed in [2], the novelty of texts is to be always determined with respect to a set of relevant information already known about the topic in concern. Thus, one cannot ascertain the novelty of a document unless s/he sees relevant source/prior information. However, the first document about a topic would always be novel if we consider a topical document stream. Novel information is often argued as an update over the existing knowledge [2]. Hence to capture this perspective, we create a target document representation that jointly encodes the target and relevant source information, which we term as the **Source Encapsulated Target Document Vector (SETDV)**. The idea is simple: we pull out the nearest source sentence corresponding to a target one and encapsulate them in one single representative sentential unit. Figure 3 shows the SETDV-CNN architecture. Here, T_1 is the *target* document whose *novelty* score is to be determined against the source document(s) $S_1, S_2, ...S_M$ i.e., to say the objective is to automatically figure out the novel information content in T_1, once the machine has already seen/scanned $S_1, S_2, ...S_M$.

1) Sentence Encoder: Instead of encoding the entire document, we encode the sentences. This makes sense as we even annotate at the sentence level. Following from [29], we train a sentence encoder on the semantically rich large-scale Stanford Natural Language Inference (SNLI) corpus and use that to generate our sentence representations for both source and target sentences. [29] show that the sentence encoder achieves the best performance with a Bi-directional LSTM followed by pooling the maximum value over each dimension of the hidden units (max pooling). We choose the training of the sentence encoder on a natural language inference (NLI) dataset because of the strong connection between textual entailment and novelty detection as previously established in the novelty subtask of RTE-TAC [15]. *If a hypothesis is inferred/entailed from a given text, it is usually non-novel. A non-entailed hypothesis, however, may contain new information*. Thus, we deem that the natural language inference task exhibits complex semantic interactions between the source (premise) and target (hypothesis) text pairs required for adjudging the novelty of the target text.

2) Encapsulation: We split the documents into constituent sentences. Then encode the sentences into their corresponding embeddings using the SNLI trained BiLSTM+max pooled sentence encoder. For each of the target sentence ***t*** we pull the most similar source sentence ***s*** using cosine similarity. We then encapsulate a target sentence with it's corresponding source as

$$ESV_t = [s, t, |s - t|, s * t]$$

where (,) signifies column vector concatenation and ESV_t is the Encapsulated Sentence Vector of a target sentence t (Figure 3). Finally, for a given target document, we stack the encapsulated sentence representations so obtained to form the Source Encapsulated Target Document Vector (SETDV) matrix. The

[9]Dataset+Code available at https://github.com/dark-archerx/To-Comprehend-the-New-On-Measuring-the-Freshness-of-a-Document

[10]we know that each event has three source documents in the corpus

[11]WMD works reasonably well for similarity search in the semantic space for shorter documents

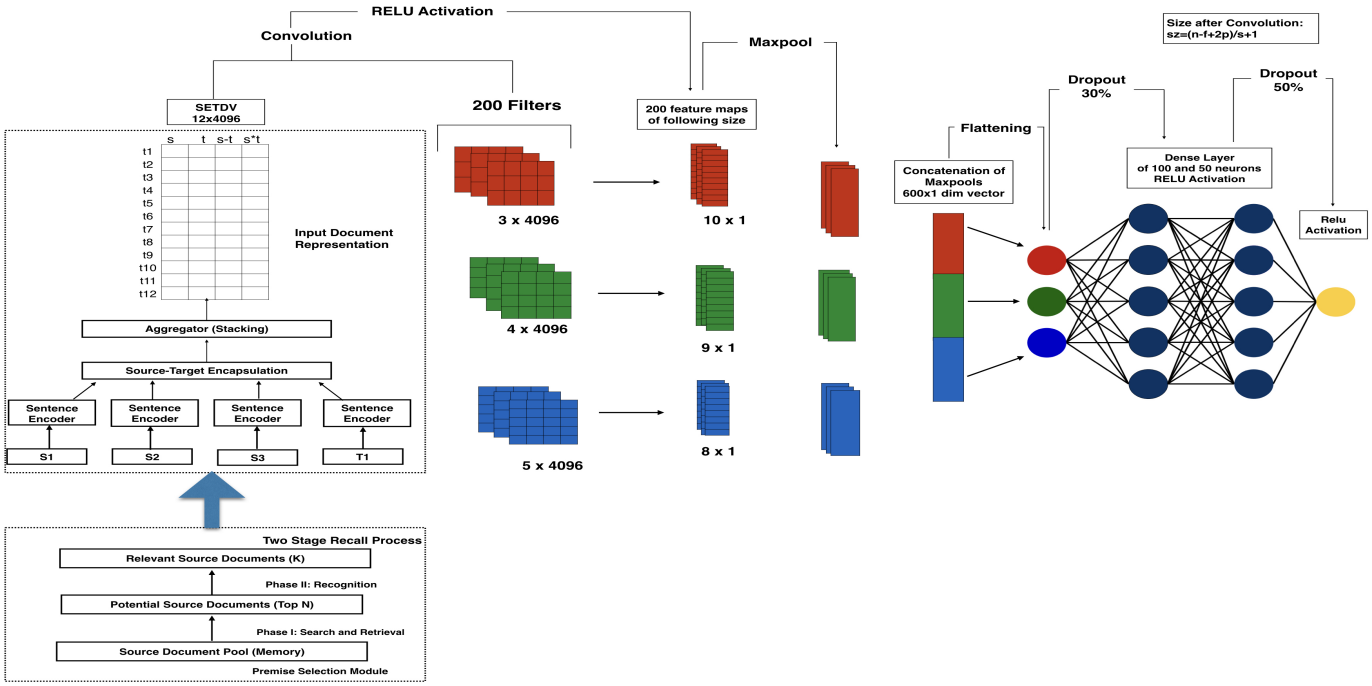

Fig. 3: The overall Novelty Score Prediction Architecture (SETDV-CNN) with 12 sentences in the target document (T1) and S1, S2, S3 $\rightarrow$ source documents

matrix has a dimension of N X $4D$ where N is the number of sentences in the target document (padded when necessary), and D is the sentence embedding dimension produced by the sentence encoder. For this representation we take inspiration from the word embedding studies by [30] where the linear offset of vectors is seen to capture semantic relationships between the two words. [31] successfully leveraged this idea for modelling sentence-pair relationships which we extend to model source-target relationships in case of documents.

Thus, we arrive at a semantic representation of the target document that has the nearest source information embedded within it. The rationale is that: a novel document would have a different semantic association with the nearest source in the vector space than that of its non-novel counterpart.

C. Convolutional Neural Network (CNN)

We use CNN as the feature extractor. Recently CNN has shown great promise in many downstream NLP applications [32]. The document matrix or the SETDV is our input to the CNN for training and subsequently predicting the novelty score of the target document with respect to the designated set of source documents. We design a CNN similar to [33] used for sentence classification. However, instead of word embeddings as input, we use the source-encapsulated target sentence embeddings of dimension $4D$ (we represent the k^{th} sentence in the document by an embedding vector $ESV_k \in \mathbb{R}^D$). We use the NON-STATIC TEXT channel variant of the CNN, where the embeddings get updated during training.

For each possible input channel, a given document is transformed into a tensor of fixed length N (padded with *zeroes* wherever necessary to tackle variable sentence lengths) by concatenating the relative sentence embeddings.

$$ESV_{1:N} = ESV_1 \oplus ESV_2 \oplus ESV_3 \oplus ..... \oplus ESV_N$$

where $\oplus$ is the concatenation operator. To extract *local features*, a convolution operation is applied. Convolution operation involves a *filter*, $W \in \mathbb{R}^{HD}$, which is convolved with a window of H embeddings to produce a local feature for the H target sentences. A local feature, c_k is generated from a window of embeddings $RSV_{k:k+H-1}$ by applying a non-linear function (*Rectified Linear Unit*) over the convoluted output. Mathematically,

$$c_k = f(W.ESV_{k:k+H-1} + b)$$

where $b \in \mathbb{R}$ is the *bias* and f is the non-linear function. This operation is applied to each possible window of H target sentences to produce a feature map (c) for the window size H.

$$c = [c_1, c_2, c_3, ..., c_{N-H+1}]$$

A global feature is then obtained by applying *max-pooling* operation [34] over the feature map. The idea behind *max-pooling* is to capture the most important feature, one with the highest value for each feature map. We describe the process by which we extract one feature from one filter (red filter portion in Figure 3 illustrate the case of $H = 3$). The model uses multiple filters for each filter size to obtain multiple features representing the text. These features form the penultimate layer, and we pass them to a fully connected feedforward network (with the number of hidden units set to 100 for the first layer and 50 for the second layer with a dropout of 0.5) followed by a *ReLU* layer whose output is the novelty prediction score.

VI. Experiments and Results

We carry on our evaluation on our dataset and automatically predict the novelty score of a document and see how it correlates with human annotated novelty score. Wherever necessary we pad the document representation with zeros.

A. Comparing Systems and Baselines

We design our baselines to serve our ablation study on the proposed model simultaneously. As comparing systems, we cover almost all published works that at any point derives a novelty score.

1) Baseline 1: We leave out SETDV and SNLI pre-training here. We take the *paragraph vector* [35] representation of the target and source documents, concatenate them and pass the joint representation through an MLP. We use the pre-trained *doc2vec* model on a newspaper corpus to generate the embeddings[12]. We select this representation as paragraph vector is known to effectively encode paragraphs/documents leveraging the power of *word2vec* [30].

2) Baseline 2: Next, we went on to investigate the importance of SNLI pre-training and implications of ablating the natural language inference knowledge for novelty detection. Textual Entailment/Natural Language Inference has been known to correlate well with the Novelty Detection [15] task. Hence, instead of taking SNLI trained semantic sentence representations, we generate them using the pre-trained *doc2vec* and use architecture identical to the proposed one (Figure 3).

3) Baseline 3: With the third baseline, we want to study how the joint encapsulation of source and target information a.k.a. **SETDV** is crucial to this task. Hence, although we generate sentence representations from pre-trained SNLI, instead of SETDV we stack the sentence representations to form the document representation. We concatenate the three source with the target document representation horizontally and feed them to the subsequent CNN module. Thus except the shape of the input matrix, this baseline resembles the proposed approach. The intuition behind each of these baselines is to let the network learn the pattern of new and redundant information only from the source and the target data representations.

4) Comparing System 1: As the first comparing system we take the popular *redundancy as opposed to novelty* technique, widely explored in several works including [17], [20], [36]. We investigate with both *tf-idf* and *doc2vec* representations of the documents. The reason being although *tf-idf* was the representation used in the original works, we also experiment with the more semantically enriched *doc2vec* to probe the actual effect of the redundancy-distance perspective to novelty scoring. We use the novelty distance metric once in pairwise *(PNov)*

$$PNov(t_i|s_1,...s_m) = min_{1\leq j\leq m}[1 - cos(t_i, s_j)] \quad (1)$$

and again in aggregate *(ANov)* form [20].

$$ANov(t_i|s_1,...s_m) = [1 - cos(t_i, S_u)] \quad (2)$$

[12]https://github.com/jhlau/doc2vec

where $S_u = \bigcup_{j=1}^{m} s_j$, t_i is the target document and s_j are the source documents.

5) Comparing System 2: We compare with the normalized blended metrics for novelty scoring introduced by [19] using cosine similarity (*cos*) and new word ratio (*nwr*) as the components.

$$J_{blended}(t_i|s_1,...s_m) = \alpha J_{nwr}(t_i) + (1-\alpha)J_{cos}(t_i) \quad (3)$$

where α is the blending parameter ranging from 0 to 1 and is learnt from our training samples ($\alpha = 0.75$).

6) Comparing System 3: We use the minimum Kullback-Leibler (KL) divergence as another comparing system [20], [36]. Thus, the respective novelty scoring formula is as follows:

$$MinKL(t_i|s_1,...s_m) = min_{1\leq j\leq m} KL(\theta_{t_i}, \theta_{s_j}) \quad (4)$$

7) Comparing System 4: Reference [21] proposed a novelty detection algorithm based on *Inverse Document Frequency* scoring function. The novelty score of a new document *d* for a collection *C* is defined as:

$$NS(d, C) = \frac{1}{norm(d)} \sum_{q \epsilon d} tf(q, d) \times idf(q, C) \quad (5)$$

where *q* is any term in target document *d*, *C* in our case are the designated source documents for *d*.

B. Results and Discussion

We deduce the novelty score of each target document with our SETDV-CNN architecture as discussed in Section V. Performance comparison of our approach with the baselines and *state-of-the-arts'* are presented in Table III. It is quite evident that SETDV-CNN is performing way better than the baselines and the comparing systems. By leveraging the power of CNN to extract features from the composite SETDV automatically, we can achieve close to human-level judgments. The reason for the low performance of the comparing systems is because those were mostly designed from an IR perspective and did not address the semantic-level information needs. However, baselines came close to the proposed approach as they manifest enriched semantic vector composition from which we extract features via neural networks.

Baseline 1 performs comparatively poor as we ablate both SNLI pre-training of the sentence vectors as well as the SETDV-CNN. In **Baseline 2** when we ablate the SNLI pre-training but keep the SETDV-CNN framework, we gain a little improvement. **Baseline 3** came out as the strongest with only the SNLI pre-training preserved. This indicates that the inference knowledge gained from training on SNLI is an important component to understand the notion of novelty. However, the higher performance of the proposed method and the adopted baselines for document-level novelty scoring clearly indicate that deep neural networks are efficient than existing feature-based and rule-based techniques for the problem under study.

Evaluation System	Description: Novelty Scoring	PC	MAE	RMSE	Cosine
Baseline 1	*doc2vec*+MLP	0.818	14.027	20.715	0.895
Baseline 2	Without SNLI pre-training	0.834	14.378	19.939	0.902
Baseline 3	Without SETDV encapsulation	0.845	13.686	18.641	0.910
Comparing System 1a	*Pairwise: tf-idf* [36], [37]	0.029	32.441	37.161	0.734
Comparing System 1b	*Pairwise: doc2vec*	0.347	40.993	54.315	0.782
Comparing System 1c	*Aggregate: tf-idf* [20]	0.130	32.281	38.901	0.728
Comparing System 1d	*Aggregate: doc2vec*	0.494	41.004	54.347	0.809
Comparing System 2a	*Blended* [38]	0.680	23.733	28.202	0.870
Comparing System 2b	*Blended using doc2vec*	0.685	40.990	54.351	0.871
Comparing System 3	Min. KLD [36]	0.592	35.997	47.718	0.846
Comparing System 4	Inverse Document Frequency [21]	0.160	41.236	54.671	0.576
Proposed Approach	**SETDV-CNN**	**0.888**	**10.294**	**16.547**	**0.953**

TABLE III: Performance of the proposed approach against the baselines and comparing systems, PC→ Pearson Correlation Coefficient, MAE→ Mean Absolute Error, RMSE→ Root Mean-Squared Error, Cosine→ Cosine simililarity between predicted and actual score vectors

We also experiment with a variant of Comparing Systems 1(b,d) and 2(b) using the semantically enriched *doc2vec* representation which supposedly gives better performance than *tf-idf* (See in Table III). It's interesting to see that our stronger baselines perform better than the *state-of-the-art's* which seconds the proposition that incorporating semantic knowledge actually improves the prediction performance. We also find that our method is more prone towards discovering redundant information, i.e. documents having low novelty score This is good considering that the dataset exhibits semantic-level redundancy. Usually, novel texts differ in lexical-level as well and hence are easier to identify. The actual challenge lies in identifying the semantically redundant textual content where the existing methods score low. Our method is efficient as we observe a higher correlation between the actual scores (human-annotated gold standard) and our system predicted scores in the scatter plot in Figure 4.

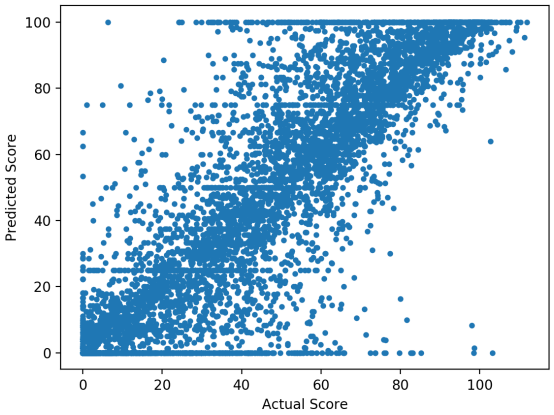

Fig. 4: Scatter plot of Actual (Gold Standard) vs Predicted (Proposed) Document-Level Novelty Score

C. Error Analysis

We analyse the predictions and identify the following class of errors committed by our system:

(1) When the target document is comparatively too small with respect to the source and other documents in the dataset. A significant amount of padding with zeros results in affinity towards non-novelty.

(2) Multiple premise scenarios: This is when a target sentence derives information from multiple source sentences. Hence, selecting only a single sentence as the source goes against our annotation perspective (during annotation we consider the overall knowledge gained from reading the source documents, not a specific source text).

(3) Target document having a complex syntactic structure as compared to the source and difficult to comprehend as well (e.g., too many complex and compound sentences).

(4) Target document has a different narrative style w.r.t. source (e.g., active vs passive voice). Such syntactic nuances were not captured correctly by our sentence encodings.

(5) Annotation conflicts among the annotators caused some errors. This happens mostly because of the (i) different novelty appetite of the annotators and (ii) not considering role of sentence significance within a document discourse for judging new information.

(6) Persistent noises, error in sentence splitting prohibited to form a complete semantic unit (Figure 2).

VII. Conclusion and Future Work

In this paper, we present probably the first deep neural method to predict the novelty score of a document. Our SETDV-CNN architecture performs close to human-annotated gold standard. The dataset we develop may pave the way for further research in understanding document-level novelty and quantifying the new information content. We believe our annotation scheme closely resembles the human understanding of new information contained in a document. It is quite unlikely that only one single source sentence would contribute towards redundancy of a target sentence. Hence dealing with multi-premise source would be our next investigation objective. Also, we would like to investigate the role of sentence significance to comprehend the novelty of an entire document for an underlying topic.

VIII. Acknowledgement

The first author and Asif Ekbal acknowledge the Visvesvaraya PhD scheme for Electronics and IT and Visvesvaraya YFRF respectively under Ministry of Electronics and Information Technology (MeitY), Government of India for support.

REFERENCES

[1] I. Soboroff and D. Harman, "Overview of the TREC 2003 novelty track," in *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003*, 2003, pp. 38–53. [Online]. Available: http://trec.nist.gov/pubs/trec12/papers/NOVELTY.OVERVIEW.pdf

[2] T. Ghosal, A. Salam, S. Tiwary, A. Ekbal, and P. Bhattacharyya, "TAP-DLND 1.0 : A corpus for document level novelty detection," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*, 2018.

[3] P. Zhao and D. L. Lee, "How much novelty is relevant? it depends on your curiosity," in *39th International ACM SIGIR Conference on Research and Development, Pisa, Italy*, 2016, p. 100.

[4] C. L. Wayne, "Topic detection and tracking (tdt)," in *Workshop held at the University of Maryland on*, vol. 27. Citeseer, 1997, p. 28.

[5] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 37–45.

[6] N. Stokes and J. Carthy, "First story detection using a composite document representation," in *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics, 2001, pp. 1–8.

[7] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 28–36.

[8] T. Brants, F. Chen, and A. Farahat, "A system for new event detection," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 330–337.

[9] I. Soboroff, "Overview of the TREC 2004 novelty track," in *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, 2004. [Online]. Available: http://trec.nist.gov/pubs/trec13/papers/NOVELTY.OVERVIEW.pdf

[10] M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, L. Zhao, and S. Ma, "Expansion-based technologies in finding relevant and new information: Thu trec 2002: Novelty track experiments," *NIST SPECIAL PUBLICATION SP*, no. 251, pp. 586–590, 2003.

[11] K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan, "Information filtering, novelty detection, and named-page finding." in *TREC*, 2002.

[12] B. Schiffman and K. R. McKeown, "Context and learning in novelty detection," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 716–723.

[13] M. Gamon, "Graph-based text representation for novelty detection," in *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 17–24.

[14] F. S. Tsai, W. Tang, and K. L. Chan, "Evaluation of novelty metrics for sentence-level novelty mining," *Information Sciences*, vol. 180, no. 12, pp. 2359–2374, 2010.

[15] L. Bentivogli, P. Clark, I. Dagan, and D. Giampiccolo, "The seventh pascal recognizing textual entailment challenge." in *TAC*, 2011.

[16] Y. Yang, J. Zhang, J. Carbonell, and C. Jin, "Topic-conditioned novelty detection," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 688–693.

[17] F. S. Tsai and Y. Zhang, "D2s: Document-to-sentence framework for novelty detection," *Knowledge and information systems*, vol. 29, no. 2, pp. 419–433, 2011.

[18] Y. Zhang and F. S. Tsai, "Combining named entities and tags for novel sentence detection," in *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval*. ACM, 2009, pp. 30–34.

[19] W. Tang, F. S. Tsai, and L. Chen, "Blended metrics for novel sentence mining," *Expert Systems with Applications*, vol. 37, no. 7, pp. 5172–5177, 2010.

[20] A. Verheij, A. Kleijn, F. Frasincar, and F. Hogenboom, "A comparison study for novelty control mechanisms applied to web news stories," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2012, pp. 431–436.

[21] M. Karkali, F. Rousseau, A. Ntoulas, and M. Vazirgiannis, "Efficient online novelty detection in news streams," in *Web Information Systems Engineering - WISE 2013 - 14th International Conference, Nanjing, China, October 13-15, 2013, Proceedings, Part I*, 2013, pp. 57–71. [Online]. Available: https://doi.org/10.1007/978-3-642-41230-1_5

[22] T. Dasgupta and L. Dey, "Automatic scoring for innovativeness of textual ideas," in *Knowledge Extraction from Text, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 12, 2016.*, 2016. [Online]. Available: http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12663

[23] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[24] E. Tulving and N. Kroll, "Novelty assessment in the brain and long-term memory encoding," *Psychonomic Bulletin & Review*, vol. 2, no. 3, pp. 387–390, 1995.

[25] M. J. Watkins and J. M. Gardiner, "An appreciation of generate-recognize theory of recall," *Journal of Memory and Language*, vol. 18, no. 6, p. 687, 1979.

[26] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004, pp. 297–304.

[27] K. W. Ng, F. S. Tsai, L. Chen, and K. C. Goh, "Novelty detection for text documents using named entity recognition," in *Information, Communications & Signal Processing, 2007 6th International Conference on*. IEEE, 2007, pp. 1–5.

[28] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015, pp. 957–966. [Online]. Available: http://jmlr.org/proceedings/papers/v37/kusnerb15.html

[29] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 2017, pp. 670–680. [Online]. Available: https://aclanthology.info/papers/D17-1070/d17-1070

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[31] L. Mou, R. Men, G. Li, Y. Xu, L. Zhang, R. Yan, and Z. Jin, "Natural language inference by tree-based convolution and heuristic matching," *arXiv preprint arXiv:1512.08422*, 2015.

[32] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[33] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1746–1751. [Online]. Available: http://aclweb.org/anthology/D/D14/D14-1181.pdf

[34] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[35] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.

[36] Y. Zhang, J. P. Callan, and T. P. Minka, "Novelty and redundancy detection in adaptive filtering," in *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, 2002, pp. 81–88. [Online]. Available: http://doi.acm.org/10.1145/564376.564393

[37] F. S. Tsai and Y. Zhang, "D2S: document-to-sentence framework for novelty detection," *Knowl. Inf. Syst.*, vol. 29, no. 2, pp. 419–433, 2011. [Online]. Available: https://doi.org/10.1007/s10115-010-0372-2

[38] F. S. Tsai and K. Luk Chan, "Redundancy and novelty mining in the business blogosphere," *The Learning Organization*, vol. 17, no. 6, pp. 490–499, 2010.