Novelty Detection: A Perspective from Natural Language Processing

Tirthankar Ghosal[*†]
Institute of Formal and
Applied Linguistics
Faculty of Mathematics and Physics
Charles University
Prague, Czech Republic
ghosal@ufal.mff.cuni.cz

Tanik Saikh
Department of Computer Science
and Engineering
Indian Institute of Technology Patna
Patna, India
1821cs08@iitp.ac.in

Tameesh Biswas
Department of Computer Science
and Engineering
Indian Institute of Technology Patna
Patna, India
biswas.cs16@iitp.ac.in

Asif Ekbal[†]
Department of Computer Science
and Engineering
Indian Institute of Technology Patna
Patna, India
asif@iitp.ac.in

Pushpak Bhattacharyya
Department of Computer Science
and Engineering
Indian Institute of Technology Bombay
Powai, India
pb@cse.iitb.ac.in

[*] The author carried out this work during his doctoral studies at the Indian Institute of Technology Patna, India.

[†] Corresponding Authors.

Submission received: 28 October 2020; revised version received: 17 October 2021; accepted fo publication: 5 December 2021.

https://doi.org/10.1162/COLI_a_00429

The quest for new information is an inborn human trait and has always been quintessential for human survival and progress. Novelty drives curiosity, which in turn drives innovation. In Natural Language Processing (NLP), Novelty Detection refers to finding text that has some new information to offer with respect to whatever is earlier seen or known. With the exponential growth of information all across the Web, there is an accompanying menace of redundancy. A considerable portion of the Web contents are duplicates, and we need efficient mechanisms to retain new information and filter out redundant information. However, detecting redundancy at the semantic level and identifying novel text is not straightforward because the text may have less lexical overlap yet convey the same information. On top of that, non-novel/redundant information in a document may have assimilated from multiple source documents, not just one. The problem surmounts when the subject of the discourse is documents, and numerous prior documents need to be processed to ascertain the novelty/non-novelty of the current one in concern. In this work, we build upon our earlier investigations for document-level novelty detection and present a comprehensive account of our efforts toward the problem. We explore the role of pre-trained Textual Entailment (TE) models to deal with multiple source contexts and present the outcome of our current investigations. We argue that a multipremise entailment task is one close approximation toward identifying semantic-level non-novelty. Our recent approach either performs comparably or achieves significant improvement over the latest reported results on several datasets and across several related tasks (paraphrasing, plagiarism, rewrite). We critically analyze our performance with respect to the existing state of the art and show the superiority and promise of our approach for future investigations. We also present our enhanced dataset TAP-DLND 2.0 and several baselines to the community for further research on document-level novelty detection.

1. Introduction

> *Of all the passions of mankind, the love of novelty most rules the mind.*
>
> *–Shelby Foote*

This quote by Shelby Foote[1] sums up the importance of novelty in our existence. Most of the breakthrough discoveries and remarkable inventions throughout history, from flint for starting a fire to self-driving cars, have something in common: They result from curiosity. A basic human attribute is the impulse to seek new information and experiences and explore novel possibilities. Humans elicit novel signals from various channels: text, sound, scene, via basic senses, and so forth. Novelty is important in our lives to drive progress, to quench our curiosity needs. Arguably the largest source of information elicitation in this digitization age is texts: be it books, the Web, papers, social media, and so forth. However, with the abundance of information comes the problem of duplicates, near-duplicates, and redundancies. Although document duplication is encouraged in certain use-cases (e.g., Content Syndication in Search Engine Optimization [SEO]), it impedes the search for new information. Hence identifying redundancies is important to seek novelties. We humans are already equipped with an implicit mechanism (*Two Stage Theory of Human Recall*: recall-recognition [Tarnow 2015]) through which we can segregate new information from old information. In our

[1] https://en.wikipedia.org/wiki/Shelby_Foote.

work, we are interested in exploring how machines would identify semantic-level non-novel information and hence pave the way to identify documents having significant content of new information. Specifically, here in this work, we investigate how we can automatically discover novel knowledge from the dimension of text or identify that a given text has new information. We rely on certain principles of Machine Learning and NLP to design efficient neural architectures for textual novelty detection at the document level.

Textual novelty detection has been known for a long time as an information retrieval problem (Soboroff and Harman 2005) where the goal is to retrieve relevant pieces of text that carry new information with respect to whatever is previously seen or known to the reader. With the exponential rise of information across the Web, the problem becomes more relevant now as information duplication (prevalence of non-novel information) is more prominent. The deluge of redundant information impedes critical, time-sensitive, and quality information to end-users. Duplicates or superfluous texts hinder reaching new information that may prove crucial to a given search. According to a particular SEO study[2] by Google in 2016, 25%–30% of documents on the Web exist as duplicates (which is quite a number!). With the emergence of humongous language models like GPT-3 (Brown et al. 2020), machines are now capable of generating artificial and semantically redundant information. Information duplication is not just restricted to lexical surface forms (mere copy), but there is duplication at the level of semantics (Bernstein and Zobel 2005). Hence, identifying whether a document contains new information in the reader's interest is a significant problem to explore to save space and time and retain the reader's attention. Novelty Detection in NLP finds application in several tasks, including text summarization (Bysani 2010), plagiarism detection (Gipp, Meuschke, and Breitinger 2014), modeling interestingness (Bhatnagar, Al-Hegami, and Kumar 2006), tracking the development of news over time (Ghosal et al. 2018b), identifying fake and misinformation (Qin et al. 2016), and so on.

As we mentioned, novelty detection as an information retrieval problem signifies retrieving relevant sentences that contain new information in discourse. Sentence-level novelty detection (Allan, Wade, and Bolivar 2003a), although important, would not suffice in the present-day deluge of Web information in the form of documents. Hence, we emphasize the problem's document-level variant, which categorizes a document (as novel, non-novel, or partially novel) based on the amount of new information in the concerned document. Sentence-level novelty detection is a well-investigated problem in information retrieval (Li and Croft 2005; Clarke et al. 2008; Soboroff and Harman 2003; Harman 2002a); however, we found that document-novelty detection attracted relatively less attention in the literature. Moreover, the research on the concerned problem encompassing semantic-level comprehension of documents is scarce, perhaps because of the argument that every document contains something new (Soboroff and Harman 2005). Comprehending the novelty of an entire document with confidence is a complex task even for humans. Robust semantic representation of documents is still an active area of research, which somewhat limits the investigation of novelty mining at the document level. Hence, categorizing a document as novel or non-novel is not straightforward and involves complex semantic phenomena of inference, relevance, diversity, relativity, and temporality, as we show in our earlier work (Ghosal et al. 2018b).

[2] https://searchengineland.com/googles-matt-cutts-25-30-of-the-webs-content-is-duplicate-content-thats-okay-180063.

This article presents a comprehensive account of the document-level novelty detection investigations that we have conducted so far (Ghosal et al. 2018b, 2019, 2021). The major contribution here is that we present our recent exploration of re-modeling multi-premise entailment for the problem and explain why it is a close approximation to identify semantic-level redundancy. We argue that to ascertain a given text's novelty, we would need multi-hop reasoning on the source texts for which we draw reference from the Question Answering (QA) literature (Yang et al. 2018). We show that our new approach achieves comparable performance to our earlier explorations, sometimes better.

We organize the rest of this article as follows: In the remainder of the current section, we motivate our current approach in light of TE. In Section 2, we discuss the related work on textual novelty detection so far, along with our earlier approaches toward the problem. Section 3 describes the current methods that utilized multiple premises for document-level novelty detection. Section 4 focuses on the dataset description. We report our evaluations in Section 5. We conclude with plans for future works in Section 6.

1.1 Textual Novelty Detection: An Entailment Perspective

TE is defined as a directional relationship between two text fragments, termed Text (T) and Hypothesis (H) as:

> *T entails H if, typically, a human reading T would infer that H is most likely true.*
> (Dagan, Glickman, and Magnini 2005).

For example, let us consider the following two texts:

Example 1

Text 1: *I left the restaurant satisfactorily*. (Premise **P**)

Text 2: *I had good food*. (Hypothesis **H**)

So a human reading Text 1 (Premise) would most likely infer that Text 2 (Hypothesis) is true, that is, Text 1 entails Text 2, or the Premise P entails the Hypothesis H.

The PASCAL-RTE challenges (Bentivogli et al. 2010, 2011) associated textual novelty with entailment. As RTE puts: *RTE systems are required to judge whether the information contained in each H is novel with respect to (i.e., not entailed by) the information contained in the corpus. If entailing sentences (T) are found for a given H, it means that the content of the H is not new (**redundant**); in contrast, if no entailing sentences are detected, it means that information contained in H is **novel***. With respect to the above example, we can say that Text 1 is known to us in a specific context. Text 2 probably has no new information to offer. However, there could be other reasons for one leaving the restaurant satisfactorily, including:

- The ambiance was good (H_1)
- The price was low (H_2)
- I got some extra fries at no cost (H_3)
- I received my birthday discount at the restaurant (H_4)

However, the probability of inferring H_1, H_2, H_3, H_4 given P seems relatively low as compared to inferring H given P in a general context.

$$Pr(H|P) > Pr(H_1|P, H_2|P, H_3|P, H_4|P)$$

Rather, we say that given P, we can implicitly assume that H is true with a higher degree of confidence. So, H might not be offering any new information. However, the same cannot be postulated for H_1, H_2, H_3, H_4 given P. Hence, the probability of H being *non-novel* given P is higher than H_1 given P, H_2 given P, H_3 given P, H_4 given P. Having said that, without a given context, H_1, H_2, H_3, H_4 are probably offering some relatively *new* information with respect to the premise P. Please note that there is a minimum lexical overlap between the Premise and the Hypothesis texts. The overlap is at the semantic level. Supposedly, TE at the semantic level is closer to detecting non-novelty.

This probabilistic nature of TE has been studied by some authors. Chen et al. (2020) introduced Uncertain Natural Language Inference (UNLI), a refinement of Natural Language Inference (NLI) that shifts away from categorical labels, targeting the direct prediction of subjective probability assessments instead. Pavlick and Kwiatkowski (2019) provide an in-depth study of disagreements in human judgments on the NLI task. They argue that NLI evaluation should explicitly incentivize models to predict distributions over human judgments. Inspired by this idea of associating entailment probabilities to texts with respect to premises, we went on to explore how we could train a machine learning architecture to identify the novelty of not only a single sentence but an entire document. However, our investigation is different from earlier explorations in the sense that:

- Novelty detection tasks in both the TREC (Soboroff and Harman 2005), and RTE-TAC (Bentivogli et al. 2011) were designed from an information retrieval perspective where the main goal was to retrieve relevant sentences to decide on the novelty of a statement. We focus on the automatic classification and scoring of a document based on its new information content from a machine learning perspective.
- As is evident from the examples, the premise-hypothesis pair shows significantly less lexical overlap, making the entailment decisions more challenging while working at the semantic level. Our methods encompass such semantic phenomena, which were less prominent in the TREC and RTE-TAC datasets.
- For ascertaining the novelty of a statement, we opine that a single premise is not enough. We would need the context, world knowledge, and reasoning over multiple facts. We discuss the same in the subsequent section.

1.2 Multiple Premise Entailment (MPE) for Novelty Detection

We deem the NLP task MPE as one close approximation to simulate the phenomenon of textual non-novelty. MPE (Lai, Bisk, and Hockenmaier 2017) is a variant of the standard TE task in which the premise text consists of multiple independently written sentences (source), all related to the same topic. The task is to decide whether the hypothesis sentence (target) can be used to describe the same topic (entailment) or cannot be used to describe the same topic (contradiction), or may or may not describe the same topic (neutral). The main challenge is to infer what happened in the topic from the multiple premise statements, in some cases aggregating information across multiple sentences into a coherent whole. The MPE task is more pragmatic than the usual TE task as it aims to assimilate information from multiple sources to decide the entailment status of the hypothesis.

Similarly, the novelty detection problem becomes more practical and hence intense when we need to consider multiple sources of knowledge (premises) to decide whether a given text (hypothesis) contains new information or not. In the real world, it is highly unlikely that a certain text would assimilate information from just another text (unlike the Premise-Hypothesis pair instances in most NLI datasets). To decide on the novelty of a text, we need to consider the context and reason over multiple facts. Let us consider the following example. Here, *source* would signify information that is already seen or known (Premise) to the reader, and *target* would signify the text for which novelty/redundancy is to be ascertained (Hypothesis).

Example 2

Source: *Survey says Facebook is still the most popular social networking site (s_1). It was created by Mark Zuckerberg and his colleagues when they were students at Harvard back in 2004 (s_2). Harvard University is located in Cambridge, Massachusetts, which is just a few miles from Boston (s_3). Zuckerberg now lives in Palo Alto, California* (s_4).

Target: *Facebook was launched in Cambridge (t_1). The founder resides in California (t_2).*

Clearly, the target text would appear *non-novel* to a reader with respect to the source/premise. However, to decide on each sentence's novelty in the target text, we would need to consider multiple sentences in the source text, not just one. Here in this case, to decide on the novelty of t_1, we would need the premises s_1, s_2, s_3 and similarly s_1, s_2, s_4 to decide for t_2. s_4 is not of interest to t_1, neither is s_3 to t_2. Thus to answer for the novelty of a certain text, it is quite likely that we may need to reason over multiple relevant sentences. Hence a multi-premise inference scenario appears to be appropriate here. In our earlier work (Ghosal et al. 2018b), we already consider *Relevance* to be one important criteria for *Novelty Detection*. So, selecting relevant premises for a statement is an important step toward detecting the novelty of the statement.

With this motivation, we design a deep neural architecture based on large-scale pre-trained TE models to find the novelty of a document. The contributions of our current work are:

- Leveraging multi-premise TE concept for document-level novelty detection with pre-trained entailment models.
- Presenting the TAP-DLND 2.0 dataset extending on TAP-DLND 1.0 (Ghosal et al. 2018b) and including sentence-level annotations to generate a document-level novelty score.

2. Related Work

In this section, we present a comprehensive discussion on the existing literature and explorations on textual novelty detection. We have been working on the document-level variant of the problem for some time. We briefly discuss our earlier approaches and learning so far before discussing our current hypothesis and approach.

2.1 Existing Literature

We survey the existing literature and advances on textual novelty detection and closely related sub-problems.

2.1.1 Early Days. Textual novelty detection has a history of earlier research (mostly from IR) with a gradual evolution via different shared tasks. We trace the first significant concern on novelty detection back to the new event/first story detection task of the Topic Detection, and Tracking (TDT) campaigns (Wayne 1997). Techniques in TDT mostly involved grouping news stories into clusters and then measuring the belongingness of an incoming story to any of the clusters based on some preset similarity threshold. If a story does not belong to any of the existing clusters, it is treated as the first story of a new event, and a new cluster is started. Vector space models, language models, lexical chain, and so forth, were used to represent each incoming news story/document. Some notable contributions in TDT are from Allan, Papka, and Lavrenko (1998); Yang et al. (2002); Stokes and Carthy (2001); Franz et al. (2001); Allan et al. (2000); Yang, Pierce, and Carbonell (1998); and Brants, Chen, and Farahat (2003). A close approximation of event-level document clustering via cross-document event tracking can be found in Bagga and Baldwin (1999).

2.1.2 Sentence-level Novelty Detection. Research on sentence-level novelty detection gained prominence in the novelty tracks of Text Retrieval Conferences (TREC) from 2002 to 2004 (Harman 2002b; Soboroff and Harman 2003; Soboroff 2004; Soboroff and Harman 2005). Given a topic and an ordered list of relevant documents, the goal of these tracks was to highlight relevant sentences that contain new information. Significant work on sentence-level novelty detection on TREC data came from Allan, Wade, and Bolivar (2003b); Kwee, Tsai, and Tang (2009); and Li and Croft (2005). Language model measures, vector space models with cosine similarity, and word count measures were the dominant approaches. Some other notable work on finding effective features to represent natural language sentences for novelty computation was based on the sets of terms (Zhang et al. 2003), term translations (Collins-Thompson et al. 2002), Named Entities (NEs) or NE patterns (Gabrilovich, Dumais, and Horvitz 2004; Zhang and Tsai 2009), Principal Component Analysis Vectors (Ru et al. 2004), Contexts (Schiffman and McKeown 2005), and Graphs (Gamon 2006). Tsai, Tang, and Chan (2010) and Tsai and Luk Chan (2010) presented an evaluation of metrics for sentence-level novelty mining.

Next came the novelty subtracks in the Recognizing Textual Entailment-Text Analytics Conferences (RTE-TAC) 6 and 7 (Bentivogli et al. 2010, 2011) where TE (Dagan et al. 2013) was viewed as one close neighbor to sentence-level novelty detection. The findings confirmed that summarization systems could exploit the TE techniques for novelty detection when deciding which sentences should be included in the update summaries.

2.1.3 Document-level Novelty Detection. At the document level, pioneering work was conducted by Yang et al. (2002) via topical classification of online document streams and then detecting novelty of documents in each topic exploiting the NEs. Zhang, Callan, and Minka (2002b) viewed novelty as an opposite characteristic to redundancy and proposed a set of five redundancy measures ranging from the set difference, geometric mean, and distributional similarity to calculate the novelty of an incoming document with respect to a set of documents in the memory. They also presented the first publicly available Associated Press-Wall Street Journal (APWSJ) news dataset for document-level novelty detection. Tsai and Zhang (2011) applied a document to sentence-level (d2s) framework to calculate the novelty of each sentence in a document that aggregates to detect novelty of the entire document. Karkali et al. (2013) computed a novelty score based on the inverse-document-frequency scoring function. Verheij et al. (2012) presented a comparative study of different novelty detection methods and evaluated them on news articles where language model-based methods performed better than the cosine similarity-based ones. More recently, Dasgupta and Dey (2016) conducted experiments with an information entropy measure to calculate the *innovativeness* of a document. Zhao and Lee (2016) proposed an intriguing idea of assessing the novelty appetite of a user based on a curiosity distribution function derived from curiosity arousal theory and the Wundt curve in psychology research.

2.1.4 Diversity and Novelty. Novelty detection is also studied in information retrieval literature for content diversity detection. The idea is to retrieve relevant yet diverse documents in response to a user query to yield better search results. Carbonell and Goldstein (1998) were the first to explore *diversity* and *relevance* for novelty with their Maximal Marginal Relevance measure. Some other notable work along this line are from Chandar and Carterette (2013) and Clarke et al. (2008, 2011). Our proposed work significantly differs from the existing literature regarding the methodology adopted and how we address the problem.

2.1.5 Retrieving Relevant Information for Novelty Detection. Selecting and retrieving relevant sentences is one core component of our current work. In recent years, there has been much research on similar sentence retrieval, especially in QA. Ahmad et al. (2019) introduced Retrieval Question Answering (ReQA), a benchmark for evaluating large-scale sentence level answer retrieval models, where they established a baseline for both traditional information retrieval (sparse term based) and neural (dense) encoding models on the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. 2016). Huang et al. (2019) explored a multitask sentence encoding model for semantic retrieval in QA systems. Du et al. (2021) introduced SentAugment, a data augmentation method that computes task-specific query embeddings from labeled data to retrieve sentences from a bank of billions of unlabeled sentences crawled from the Web. Yang et al. (2020) uses the Universal Sentence Encoder (USE) for semantic similarity and semantic retrieval in a multilingual setting. However, in our current work, we apply a simple TE probability-based ranking method to rank the relevant source sentences with respect to a given target query sentence.

2.2 Our Explorations So Far

As is evident from our discussion so far, textual novelty detection was primarily investigated in the Information Retrieval community, and the focus was on novel sentence retrieval. We began our exploration on textual novelty detection with the motivation to cast the problem as a document classification task in machine learning. The first hurdle we came across was the non-availability of a proper document-level novelty detection dataset that could cater to our machine learning experimental needs. We could refer to the only available dataset, the APWSJ (Zhang, Callan, and Minka 2002a). However, APWSJ too was not developed from a machine learning perspective as the dataset is skewed toward *novel* documents (only 8.9% instances are *non-novel*). Hence, we decided to develop a dataset (Ghosal et al. 2018b) from newspaper articles. We discuss our dataset in detail in Section 4.1. Initially, we performed some pilot experiments to understand the role of TE in textual novelty detection (Saikh et al. 2017). We extracted features from source-target documents and experimented with several machine learning methods, including Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF), and so on. We also investigated our idea of TE-based novelty detection on the sentence-level entailment-based benchmark datasets from the Recognizing Textual Entailment (RTE) tasks (Bentivogli et al. 2010, 2011).

We discuss the approaches we developed so far in the subsequent section.

2.2.1 Feature-based Method for Document Novelty Detection. We view novelty as an opposite characteristic to Semantic Textual Similarity (STS), with our first investigation (Ghosal et al. 2018b) on document-level novelty detection as a classification problem. We curate several features from a target document (with respect to a predefined set of source documents) like *paragraph vector (doc2vec) similarity, KL divergence, summarization similarity (concept centrality using TextRank [Mihalcea and Tarau 2004]), lexical n-gram similarity, new words count, NE and keyword similarity,* and so forth, and build our classifier based on RF. The dominant feature for the classification was *new word count* followed by *document-level semantic similarity, keyword, and named-entity similarity*.

2.2.2 RDV-CNN Method for Document Novelty. Next we develop a deep neural architecture (Ghosal et al. 2018a) to classify documents as *novel* or *non-novel* based on *new* information content. We represent our target documents as semantic vectors. We train our sentence encoders on the semantically rich, large-scale (570k sentence pairs) Stanford Natural Language Inference (SNLI) dataset (Bowman et al. 2015). We generate sentence encodings by feeding GloVe word vectors to a Bi-Directional LSTM followed by max pooling (Conneau et al. 2017). We arrive at a certain document level semantic representation (inspired from Mou et al. [2016]) that models both source and target information in a single entity, which we term the **Relative Document Vector (RDV)**. Each sentence in the target document is represented as:

$$RSV_k = [a_k, b_{ij}, |a_k - b_{ij}|, a_k * b_{ij}]$$

where RSV_k is the **Relative Sentence Vector (RSV)** of sentence k in the target document, a_k is the sentence embedding of the target sentence k, and b_{ij} is the sentence embedding of the i-th sentence in source document j. We selected the nearest premise source sentence ij using cosine similarity. We stack the RSV corresponding to all the target sentences to form the RDV. The RDV becomes the input to a deep Convolutional Neural Network (CNN) (Kim 2014) for automatic feature extraction and subsequent classification of a document as *novel* or *non-novel*. We extend this idea to compute the document-level novelty score in Ghosal et al. (2019).

2.2.3 Detecting Document Novelty via Decomposable Attention. With our subsequent investigation (Ghosal et al. 2021) we experiment with a decomposable attention-based deep neural approach inspired by Bowman et al. (2015) and Parikh et al. (2016). For a semantically redundant document (*non-novel*), we contend that the neural attention mechanism would be able to identify the sentences in the source document that has identical information and is responsible for *non-novelty* of the target document (we call it **Premise Selection**). We then jointly encode the source-target alignment and pass it through an MLP for classification. This approach is simple with an order of fewer parameters as compared to other complex deep neural architectures. Inspired by works on attention in the Machine Translation literature (Bahdanau, Cho, and Bengio), it relies only on the learning of sentence-level alignments to generate document-level novelty judgments.

Our current work differs from the existing literature on novelty detection, even from our earlier attempts in many aspects. The majority of earlier prominent work on novelty detection focused on novel sentence retrieval. In our earlier attempts, we did not consider multiple premises for ascertaining the novelty of an information unit (sentence in our case). Here, we attempt a *multi-hop multi-premise entailment* to address the scenario we discussed in Section 1.2. Assimilating information from multiple sources and enhancing the retrieved source information with their relevant weights are some crucial contributions for document-level novelty detection in this work. Finally, we introduce a novel dataset to quantify document novelty.

A somewhat similar work for information assimilation from multiple premises is Augenstein et al. (2019) where the authors perform automatic claim verification from multiple information sources. In that work, the authors collect claims from 26 fact-checking Web sites in English, pair them with textual sources and rich metadata, and label them for veracity by human expert journalists. Although our work encompasses information assimilation from multiple sources, we differ from Augenstein et al. (2019) in the motivation and the task definitions. However, we can draw parallels with our work as novel facts would be hard to verify because there would not be enough evidence to corroborate those facts' claims. However, if a fact is entailed from authentic information sources, it can be verified, which means that it would not be a novel one. The discussion opens up an interesting aspect: A verified fact contains information that could be entailed from authentic information sources; hence the fact would not be saying something drastically new. A fact that is novel would be hard to verify due to a lack of prior information.

3. Current Methodology: Encompassing Multiple Premises for Document-Level Novelty Detection

As discussed in Section 1.2, reasoning over multiple facts is essential for textual novelty detection. We may need to assimilate information from multiple source texts to ascertain the state of the novelty of a given statement or a fact. *If a text is redundant against a given prior, it is redundant against the set of all the relevant priors. However, it has to be novel against all the relevant priors for a text to be novel.* Here, a prior signifies the relevant information exposed to the reader that s/he should refer to determine the *newness* of the *target text*. If no such priors are available, possibly the target text has new information. Organizers of TREC information retrieval exercises (Soboroff 2004) formulated the tasks along this line. If for a given query (target), no relevant source is found from a test collection, possibly the query is new. Here s_1, s_2, s_3, s_4 are the relevant priors for t_1, t_2.

We also indicate in our earlier work (Ghosal et al. 2019) that the selection of relevant prior information is an essential precursor toward deciding the novelty of a given statement or fact. Hence, finding the relevant source sentences is essential toward ascertaining the *newness* of the target sentence. Hence, in our proposed approach, we encompass two components:

- a relevance detection module, followed by
- a novelty detection module

We make use of pre-trained NLI models for both the components. To assimilate information from multiple priors, the novelty detection module manifests a **join** operation at multiple layers of the pre-trained entailment stack to capture multiple levels of abstraction. The join operation is inspired by Trivedi et al. (2019) for QA. It results in a multi-premise (source) aware hypothesis (target) representation, where we combine all such target sentence representations to decide on the novelty of the target document. Figure 1a shows the architecture of our proposed approach.

3.1 Relevance Detection

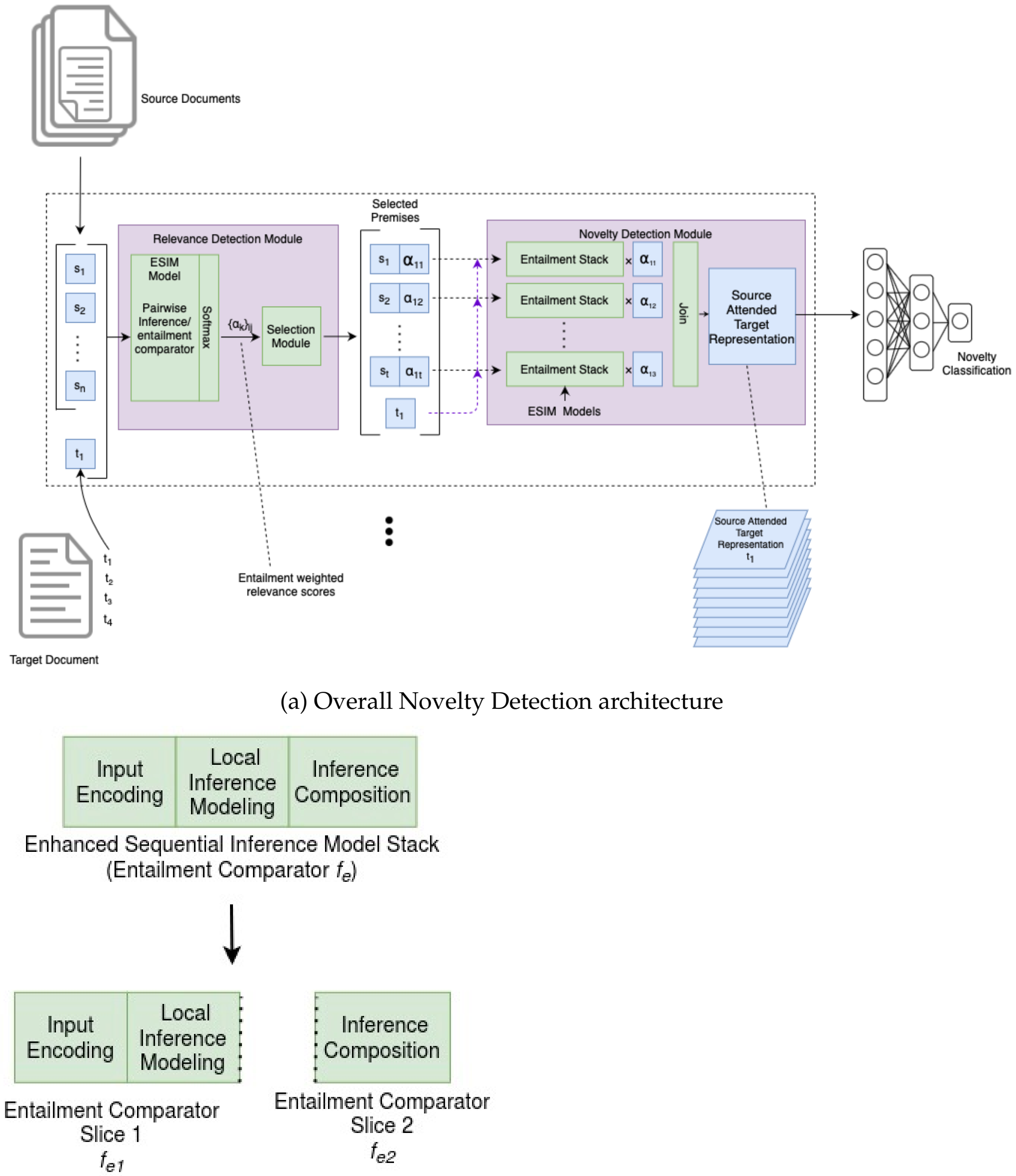
The goal of this module is to find relevant premises (source sentences) for each sentence in the target document. We treat the sentences in the target document as our multiple hypotheses, that is, we understand a target document to comprise multiple hypothesis statements. The objective is to find to what extent each of these hypotheses is entailed from the premises in the source documents and use that knowledge to decide the target document's novelty. Ideally, *a non-novel document would find the majority of its sentences highly entailed from the various sentences in the source documents*. A source sentence is considered relevant if it contains information related to the target sentence and may serve as the premise to determine the newness of the target sentence. We model this relevance in terms of entailment probabilities, that is, how well the information in the source and the target correlate. We use a pre-trained inference model to give us the entailment probabilities between all possible pairs of target and source sentences. Not all sentences in the source documents would be relevant for a given target sentence (as per the example in Section 1.2, s_4 is not relevant for t_1 and s_3 is not relevant to t_2). For each target sentence (t_k), we select the top f source sentences with the highest entailment probabilities (α_{kf}) as the relevant priors. After softmax, the final layer of a pre-trained entailment model would give us the entailment probability between a given premise-hypothesis pair.

3.1.1 Input. Let S_1, S_2, ..., S_n be the source documents retrieved from a document collection for a target document T. In our experiments, we already had the source documents designated for a given target document. We split the source and target documents into corresponding sentences. Here, s_{ij} denotes the ith sentence of the source document j. t_k represents the sentences in the target document (T). The final objective is to determine whether T is *novel* or *non-novel* with respect to S_1, S_2, ..., S_n.

3.1.2 Inference Model. The source-target sentence pairs are then fed to a pre-trained NLI model to obtain the entailment probabilities after the final (softmax activation) layer. Here, we make use of the Enhanced Sequential Inference Model (ESIM) (Chen et al. 2017) trained on large-scale inference datasets, SNLI (Bowman et al. 2015) and MultiNLI (Williams, Nangia, and Bowman 2018), as our pre-trained entailment stack.

$$\{\alpha_k\}_{ij} := Pr[s_{ij} \rightarrow t_k]$$

where $\{\alpha_k\}_{ij}$ denotes probability of entailing t_k from source sentence s_{ij}. This is the output of the pre-trained ESIM model's softmax layer on Premise s_{ij} and Hypothesis t_k.


Source Documents
Relevance Detection Module
ESIM Model
Pairwise Inference/ entailment comparator
Softmax
$\{\alpha_k\}_i$
Selection Module
s_1
s_2
s_n
t_1
Selected Premises
s_1 α_{11}
s_2 α_{12}
s_t α_{1t}
t_1
Novelty Detection Module
Entailment Stack × α_{11}
Entailment Stack × α_{12}
Entailment Stack × α_{13}
ESIM Models
Join
Source Attended Target Representation
Novelty Classification
Target Document
t_1
t_2
t_3
t_4
Entailment weighted relevance scores
Source Attended Target Representation t_1

(a) Overall Novelty Detection architecture

Input Encoding
Local Inference Modeling
Inference Composition
Enhanced Sequential Inference Model Stack (Entailment Comparator f_e)
Input Encoding
Local Inference Modeling
Entailment Comparator Slice 1 f_{e1}
Inference Composition
Entailment Comparator Slice 2 f_{e2}

(b) The usual ESIM entailment stack sliced for Local Inference and Global Inference Composition

Figure 1
Multi-premise entailment-based document-level novelty detection architectures' overview. It has two components: the Relevance Detection module, which computes relevance scores, and the Novelty Detection module, which aggregates multiple premises, computes entailment, and classifies the target document. The entailment model in the relevance module uses full entailment stack (ESIM in this case), whereas the novelty module uses multiple partial entailment stacks (excluding the last projection layer) to aggregate the premises via a join operation.

Predicting subjective entailment probabilities instead of inference categories is explored by Chen et al. (2020), where they use the term *Uncertain NLI*.

3.2 Selection Module and Relevance Scores

Not all the source sentences would contribute toward the target sentence. Hence, we retain the topmost f relevant source sentences for the target sentence t_k based on the entailment probabilities or what we term as the **relevance scores**. In Figure 1, α_{kf} denotes the relevance scores for the top f selected source sentences for a target sentence t_k. We would further use these relevance scores while arriving at a Source-Aware Target (SAT) representation in the Novelty Detection module. Thus, the relevance module's outputs are multiple relevant source sentences s_{kf} for a given target sentence t_k and their pairwise relevance scores.

3.3 Novelty Detection Module

The goal of the Novelty Detection module is to assimilate information from the multiple relevant source sentences (from source documents) to ascertain the novelty of the target document. The novelty detection module would take as input the target document sentences paired with their corresponding f relevant source sentences. This module would again make use of a pre-trained entailment model (i.e., ESIM here) along with the relevance scores between each source-target sentence pair from the earlier module to independently arrive at a SAT representation for each target sentence t_k. We use the earlier module's relevance scores to incentivize the contributing source sentences and penalize the less-relevant ones for the concerned target sentence. Finally, we concatenate the k SAT representations, passing it through a final feed-forward and linear layer, to decide on the novelty of T. We discuss the assimilation of multiple premises weighted by their relevance scores in the following section. The number of entailment functions in this layer depends on the number of target sentences (k) and the number of relevant source sentences you want to retain for each target sentence (i.e., f).

3.3.1 Relevance-weighted Inference Model to Support Multi-premise Entailment. A typical neural entailment model consists of an input encoding layer, local inference layer, and inference composition layer (see Figure 1b). The input layer encodes the premise (source) and hypothesis (target) texts; the local inference layer makes use of cross-attention between the premise and hypothesis representations to yield entailment relations, followed by additional layers that use this cross-attention to generate premise attended representations of the hypothesis and vice versa. The final layers are classification layers, which determine entailment based on the representations from the previous layer. In order to assimilate information from multiple source sentences, we use the *relevance scores* from the previous module to scale up the representations from the various layers of the pre-trained entailment model (E) and apply a suitable join operation (Trivedi et al. 2019). In this join operation, we use a part of the entailment stack to give us a representation for each sentence pair that represents important features of the sentence pair and hence gives us a meaningful document level representation when combined with weights. We denote this part of the stack as f_{e1}. The rest of the entailment stack that we left out in the previous step is used to obtain the final representation from the combined intermediate representations and is denoted by f_{e2}. This way, we aim to emphasize the top relevant source-target pairs and attach lesser relevance scores to the bottom ones for a given target sentence t_k. The join operation would facilitate the assimilation of multiple source information to infer on the target.

We now discuss how we incorporate the relevance scores to various layers of the pre-trained entailment model (E) and assimilate the multiple source information for a given target sentence t_k.

3.3.2 Input Layer to Entailment Model. For convenience, let us denote any source sentence (premise) as **s** and any target sentence (hypothesis) as **t**.

$$s = (x_1, x_2, x_3, ...x_{l_s})$$
$$t = (y_1, y_2, y_3, ...y_{l_t})$$

where $x_1, x_2, x_3, ...$ are tokens of source sentence **s** and $y_1, y_2, y_3, ...$ are tokens of target sentence **t**. The length of **s** and **t** are l_s and l_t, respectively.

There is a BiLSTM encoder to get the representation of **s** and **t** as:

$$\bar{s}_i = \{BiLSTM(s)\}_i, i \in (1, 2, ...l_s)$$
$$\bar{t}_j = \{BiLSTM(t)\}_j, j \in (1, 2, ...l_t)$$

where $\bar{s}_i$ denotes the output vector of BiLSTM at the position i of the premise, which encodes word s_i and its context.

3.3.3 Cross-Attention Layer. Next is the cross attention between the source and target sentences to yield the entailment relationships. In order to put emphasis on the most relevant source-target pairs, we scale the cross-attention matrices with the relevance scores from the previous module and then re-normalize the final matrix.

Cross-attention between source to target and target to source is defined as:

$$\tilde{s}_i = \sum_{j=1}^{l_t} \frac{exp(e_{ij})}{\sum_{k=1}^{l_t} exp(e_{ik})} \bar{t}_j$$
$$\tilde{t}_j = \sum_{i=1}^{l_s} \frac{exp(e_{ij})}{\sum_{k=1}^{l_s} exp(e_{jk})} \bar{a}_i$$

where, $e_{ij} = (\bar{s}_i)^T \bar{t}_j$.

So, for a source sentence s against a given target sentence t, we obtain a source to target cross-attention matrix $\tilde{s}_i$ and a target to source cross-attention matrix $\tilde{t}_j$ with dimension ($i \times j$) and ($j \times i$), respectively.

Now for our current multi-source and multi-target scenario, for the given target sentence t_k, we found f relevant source sentences $s_{k1}, s_{k2}, \ldots, s_{kf}$. The assimilation mechanism would scale the corresponding attention matrices by a factor α_{kf} for each source (s_f)-target (t_k) pair to generate the SAT for t_k against $s_{k1}, s_{k2}, \ldots, s_{kf}$.

We scale the cross-attention matrix with the *relevance scores* (α_{kf}) to prioritize the important source sentences for a given target sentence and concatenate the matrices for all the f source sentences ($s_{k1}, s_{k2}, \ldots, s_{kf}$) against a given target sentence t_k.

$$\tilde{s}_i^{s_{kf}t_k} = [\alpha_{k1}\tilde{s}_i^{s_{k1}t_k}; ....; \alpha_{kf}\tilde{s}_i^{s_{kf}t_k}]$$

where k remains unchanged for a given t_k and f varies for the multiple source sentences against a given t_k.

We concatenate the source sentences $(s_{k1}, s_{k2}, \ldots., s_{kf})$ for a given t_k to obtain the passage-level representation as:

$$[S_{kf}] = [[\alpha_{k1}\bar{s}_{k1}]; [\alpha_{k2}\bar{s}_{k2}]; ....; [\alpha_{kf}\bar{s}_{kf}]]$$

We keep the target sentence representation $(\bar{t}_k)$ unchanged. We forward the scaled attention matrices, scaled source representations, and the unchanged target representation to the next layer in the entailment stack. We repeat the same operation for all the sentences $(t_1, t_2, \ldots, t_k)$ in the target document T.

3.3.4 Source-Aware Target Representations. We also scale the final layer in the entailment stack (E_{kf}) with the *relevance scores* (α_{kf}). The final layer in the entailment stack usually outputs a single vector $\bar{h}$, which is then used in a linear layer and a final logit to obtain the final decision. Here, the join operation is a weighted sum of the source-target representations from the preceding layers. So we have:

$$SAT_k = \sum_{f} \alpha_{kf} h_{kf}$$

where SAT_k is the Source-Aware Target representation for t_k. We do the same for all the target sentences in the target document T.

Selected source premises (s_{kf}) from the selection module are scaled with the relevance attention weights (α_{kf}) to attach importance to the selected premises. The transformation from s_{kf} to h_{kf} is achieved by cross-attention between source and target sentences followed by a concatenation of the attention-weighted premise, followed by the higher-order entailment layers in the ESIM stack (pooling, concatenation, feed-forward, linear) (Chen et al. 2017). h_{kf} is the output of the entailment stack, which is further scaled with the attention weight (α_{kf}). For further details on how the ESIM stack for inference works (for e.g., the transformation of source representations to the entailment hidden state representations), please consult Chen et al. (2017).

3.3.5 Novelty Classification. We stack the SAT representations (SAT_k) for all the sentences in the target document and pass the fused representation through an MLP to discover important features and finally classify with a layer having softmax activation function. The output is whether the target document is *Novel* or *Non-Novel* with respect to the source documents.

4. Dataset Description

The most popular datasets for textual novelty detection are the ones released in TREC 2002–2004 (Harman 2002a; Soboroff and Harman 2003) and RTE-TAC 2010–2011 (Bentivogli et al. 2010, 2011). However, these datasets are for sentence-level novelty mining and hence do not cater to our document-level investigation needs. Therefore, for the current problem of the document-level novelty classification, we experiment with two document-level datasets: the APWSJ (Zhang, Callan, and Minka 2002b), and the one we developed—TAP-DLND 1.0 (Ghosal et al. 2018b). We also extend our TAP-DLND 1.0 dataset, include sentence-level annotations to arrive at a document-level novelty score, and coin it as **TAP-DLND 2.0**, which we present in this article. All these datasets are in the newswire domain.

4.1 TAP-DLND 1.0 Corpus

We experiment with our benchmark resource for document-level novelty detection (Ghosal et al. 2018b). The dataset is balanced and consists of 2,736 *novel* and 2,704 *non-novel* documents. There are several categories of events; ten to be precise (Business, Politics, Sports, Arts and Entertainment, Accidents, Society, Crime, Nature, Terror, Society). For each novel/non-novel document, there are three source documents against which the target documents are annotated. While developing this dataset, we ensured that *Relevance, Relativity, Diversity, and Temporality* (Ghosal et al. 2018b) characteristics were preserved.

For developing this resource, we tracked the development of an event (news items) across time over several Indian newspapers. We did a temporal crawling of event-specific news items published by different newspapers over a specific period. For a particular event, we select a set of documents as the *source* knowledge or the *prior relevant knowledge* of the event and the rest as *target* documents (for which the state of novelty would be ascertained). The core idea is: For a given event (*e.g., reporting of an accident in Bali*), the different newspapers would report more or less similar content on a given date. On the subsequent dates, new information regarding the event may surface up (e.g., *the accident was actually a plot*). The relevant temporal information update over the existing knowledge is what we deem as *novel knowledge*. We intentionally chose such events, which continued in the news for some days to facilitate our notion of novelty update. We ask our annotators to judge the target document's information against the source documents only [Annotation Label: NOVEL or NON-NOVEL]. We follow the following **annotation principles**:

1. To annotate a document as *non-novel* whose semantic content significantly overlaps with the source document(s) (maximum redundant information).
2. To annotate a document as *novel* if its semantic content, as well as intent (direction of reporting), significantly differs from the source document(s) (minimum or no information overlap). It could be an update on the same event or describing a post-event situation.
3. We left out the ambiguous cases (for which the human annotators were unsure about the label).

Our dataset manifests the presence of semantic-level redundancies, goes beyond lexical similarity, and hence it makes an ideal candidate for our experiments. With respect to the chosen source documents, we found novel documents appearing in later dates of the event in chronological order, and the non-novel documents are found from the initial days of the event reporting (usually the dates from which the source documents are selected). The inter-rater agreement is 0.82 in terms of the Fleiss Kappa (Fleiss 1971), and the average length of documents is 15 sentences/353 words. Figure 2 shows the organization of our dataset.

Apart from the inter-rater agreement, we use Jaccard Similarity (Jaccard 1901), BLEU (Papineni et al. 2002), and ROUGE (Lin 2004) to judge the quality of data. We compute the average scores between source and target documents and show this in

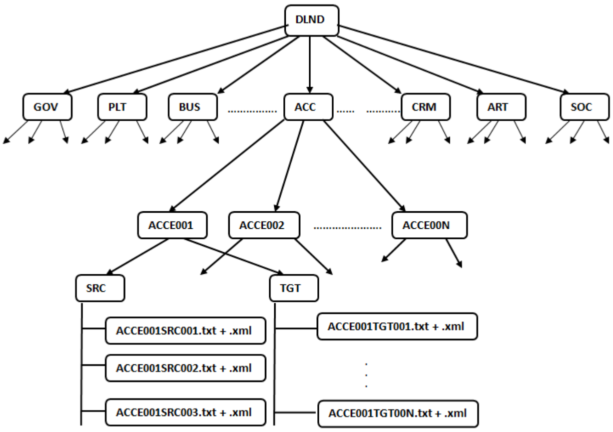


Figure 2
The TAP-DLND 1.0 corpus structure. We retain the structure in the extended dataset (TAP-DLND 2.0) we use in the current work.

Table 1
On measuring quality of annotations via automatic metrics (weak indicators).

Metrics	Novel	Non-Novel
Jaccard Similarity	0.069	0.134
BLEU	0.055	0.193
ROUGE	0.281	0.408

Table 1. It is clear that non-novel documents' similarity with the corresponding source documents are higher compared to their novel counterparts, which is justified.

4.2 APWSJ Dataset

The APWSJ dataset consists of news articles from the Associated Press (AP) and Wall Street Journal (WSJ) covering the same period (1988–1990) with many on the same topics, guaranteeing some redundancy in the document stream. There are 11,896 documents on 50 topics (Q101–Q150 TREC topics). After sentence segmentation, these documents have 319,616 sentences in all. The APWSJ data contain a total of 10,839 (91.1%) novel documents and 1,057 (8.9%) non-novel documents. However, similar to Zhang, Callan, and Minka (2002b), we use the documents within the designated 33 topics with redundancy judgments by the assessors. The dataset was meant to filter superfluous documents in a retrieval scenario to deliver only the documents having a redundancy score below a calculated threshold. Documents for each topic were delivered chronologically, and the assessors provided two degrees of judgments on the non-novel documents: *absolute redundant* or *somewhat redundant*, based on the preceding documents. The unmarked documents are treated as *novel*. However, because there is a huge class imbalance, we follow Zhang, Callan, and Minka (2002b), and include the somewhat redundant documents also as *non-novel* and finally arrive at ∼37% non-novel instances. Finally, there are 5,789 total instances, with 3,656 novel and 2,133 non-novel. The proportion of novel instances for the novelty classification experiments is 63.15%.

4.3 TAP-DLND 2.0 Corpus

We present the extended version of our TAP-DLND 1.0 corpus with this work. The new TAP-DLND 2.0 dataset is available at `https://github.com/Tirthankar-Ghosal/multipremise-novelty-detection`. Whereas TAP-DLND 1.0 is for document-level novelty classification, the TAP-DLND 2.0 dataset is catered toward deducing the novelty score of a document (quantifying novelty) based on the information contained in the preceding/source documents. Also, we annotate the new dataset at the sentence level (more fine-grained) in an attempt to weed out inconsistencies that may have persisted with document-level annotations.

We re-annotate TAP-DLND 1.0 from scratch, now at the sentence level, extend to more than 7,500 documents, and finally deduce a document-level novelty score for each target document. The judgment of novelty at the document level is not always unanimous and is subjective. Novelty comprehension also depends on the appetite of the observer/reader (in our case, the annotator or the labeler) (Zhao and Lee 2016). It is also quite likely that every document may contain something new with respect to previously seen information (Soboroff and Harman 2003). However, this relative amount of new information is not always justified to label the entire document as novel. Also, the significance of the new information with respect to the context plays a part. It may happen that a single information update is so crucial and central to the context that it may affect the novelty comprehension of the entire document for a labeler. Hence, to reduce inconsistencies, we take an objective view and deem that instead of looking at the target document in its entirety, if we look into the sentential information content, we may get more fine-grained new information content in the target document discourse. Thus, with this motivation, we formulate a new set of annotation guidelines for annotations at the sentence level. We associate scores with each annotation judgment, which finally cumulates to a document-level novelty score. We design an easy-to-use interface (Figure 4) to facilitate the annotations and perform the annotation event-wise. For a particular event, an annotator reads the predetermined three seed source documents, gathers information regarding that particular event, and then proceeds to annotate the target documents, one at a time. Upon selecting the desired target document, the interface splits the document into constituent sentences and allows six different annotation options for each target sentence (cf. Table 2). 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. The current data statistics for TAP-DLND 2.0 is in Table 3. We also plot the correspondence between the classes of TAP-DLND 1.0 and the novelty scores of TAP-DLND 2.0 to see how the perception of novelty varied across sentence and document-level annotations. The plot is in Figure 3. We divide the whole range of novelty scores (from TAP-DLND 2.0 annotations) within a set of five intervals, which are placed in the x-axis. The number of novel/non-novel annotated documents (from TAP-DLND 1.0) are shown in the vertical bars. We can see that the number of novel documents steadily increases as the novelty score range increases, while the reverse scenario is true for non-novel documents. This behavior signifies that the perception did not change drastically when we moved from document-level to sentence-level annotations and also that our assigned scores (in Table 2) reflect this phenomena to some extent.

Table 2
Sentence-level annotations. The target document sentences are annotated with respect to the information contained in the source documents for each event. The annotations are qualitatively defined. We assign scores to quantify them.

Annotation Labels	Description	Score
Novel (NOV)	The entire sentence has new information.	1.00
Non-Novel (NN)	The information contained in the sentence is completely redundant.	0.00
Mostly Non-Novel (PN25)	Most of the information overlaps with the source with little new information.	0.25
Partially 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 3
TAP-DLND 2.0 dataset statistics. Inter-rater agreement (Fleiss 1971) is measured for 100 documents for sentence-level annotations by two raters.

Dataset Characteristics	Statistics
Event categories	10
Number of events	245
Number of source documents per event	3
Total target documents	7,536
Total sentences annotated	**120,116**
Average number of sentences per document	~ 16
Average number of words per document	~ 320
Inter-rater agreement	0.88

4.3.1 About the Annotators. We had the same annotators from TAP-DLND 1.0 working on the TAP-DLND 2.0 dataset. One of the two full-time annotators holds a master's degree in Linguistics, and the other annotator holds a master's degree in English. They were hired full-time and paid the usual research fellow stipend in India. The third annotator to resolve the differences in the annotations is the first author of this article. The annotation period lasted more than six months. On average, it took $\sim$30 minutes to annotate one document of average length, but the time decreased and the consensus increased as we progressed in the project. A good amount of time went into reading the source documents carefully and then proceeding toward annotating the target document based on the acquired knowledge from the source documents for a given event. Because the annotators were already familiar with the events and documents (as they also did the document-level annotations for TAP-DLND 1.0), it was an advantage for them to do the sentence-level annotations.

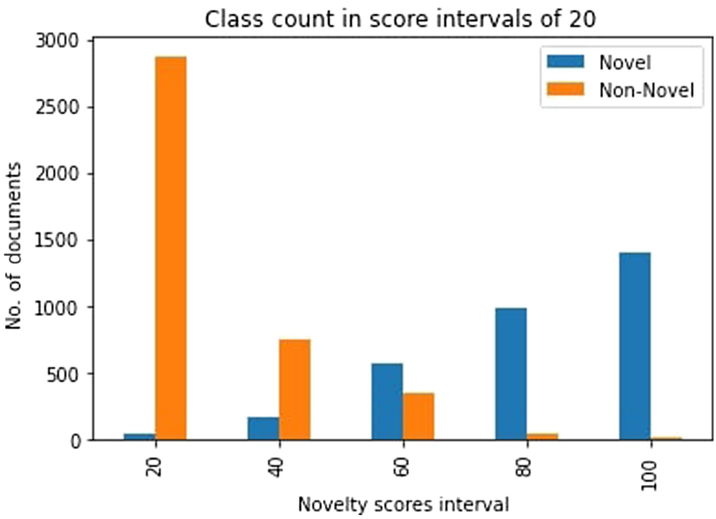
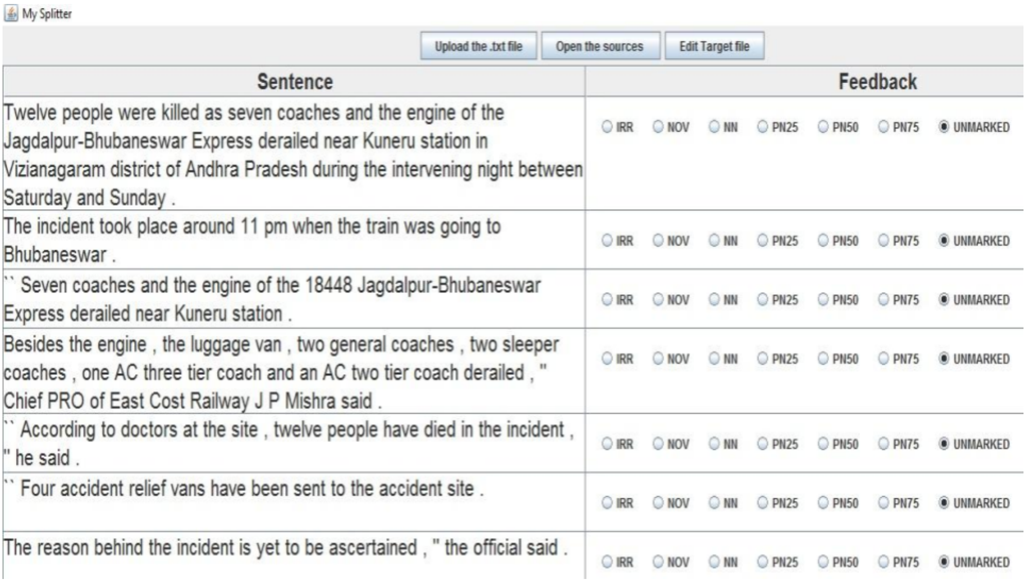


Figure 3
The novelty class and novelty score correspondence between TAP-DLND 1.0 and TAP-DLND 2.0 datasets. The blue bars and orange bars represent number of novel and non-novel documents (y-axis) in the given score range (x-axis), respectively.



Sentence	Feedback
Twelve people were killed as seven coaches and the engine of the Jagdalpur-Bhubaneswar Express derailed near Kuneru station in Vizianagaram district of Andhra Pradesh during the intervening night between Saturday and Sunday .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED
The incident took place around 11 pm when the train was going to Bhubaneswar .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED
`` Seven coaches and the engine of the 18448 Jagdalpur-Bhubaneswar Express derailed near Kuneru station .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED
Besides the engine , the luggage van , two general coaches , two sleeper coaches , one AC three tier coach and an AC two tier coach derailed , " Chief PRO of East Cost Railway J P Mishra said .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED
`` According to doctors at the site , twelve people have died in the incident , " he said .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED
`` Four accident relief vans have been sent to the accident site .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED
The reason behind the incident is yet to be ascertained , " the official said .	○ IRR ○ NOV ○ NN ○ PN25 ○ PN50 ○ PN75 ● UNMARKED

Figure 4
The sentence-level annotation interface used to generate the document-level novelty score (gold standard).

4.3.2 Annotation Example. 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 with respect to the source text. It is quite clear that the target text has new information with respect to the source, except that the first sentence in the target contains some redundant content (*Singapore is a city-state*). Analyzing the first sentence in the target text, we obtain two pieces of 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, that is, 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 with respect to the source text. If all the sentences were tagged as novel, the score would have been 3.0, indicating 100% novel information in the target text. So, if a target document *X* has *n* sentences, and the novelty annotation for each sentence is n_i, the document-level novelty-score of X would be:

$$[n_1 + n_2 + ...n_n]/n$$

where n_i can assume the values from Table 2.

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, 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. Identifying and annotating an information unit would be complex. However, we plan for further research with annotation at the phrase-level and with relative importance scores.

4.4 Datasets for Allied Tasks

Finding semantic-level redundancy is more challenging than finding novelty in texts (Ghosal et al. 2018a). The challenge scales up when it is at the level of documents. Semantic-level redundancy is a good approximation of non-novelty. Novel texts usually consist of new terms and generally are lexically different from the source texts. Hence with our experiments, we stress on detecting non-novelties, which would eventually lead us to identify novelties in text. Certain tasks could simulate the detection of non-novelty. Paraphrasing is one such linguistic task where paraphrases convey the same information as the source texts yet have a significantly less lexical similarity. Another task that comes close to identifying novelties in the text is plagiarism detection, which is a common problem in academia. We train our model with the document-level novelty datasets and test its efficacy to detect paraphrases and plagiarized texts. We use the following well-known datasets for our investigation.

4.4.1 Webis Crowd Paraphrase Corpus. The Webis Crowd Paraphrase Corpus 2011 (Webis-CPC-11) (Burrows, Potthast, and Stein 2013) consists of 7,859 candidate paraphrases obtained from the Amazon Mechanical Turk crowdsourcing. The corpus[3] is made up of 4,067 accepted paraphrases, 3,792 rejected non-paraphrases, and the original texts. For our experiment, we assume the original text as the source document and the corresponding candidate paraphrase/non-paraphrase as the target document. We hypothesize that a paraphrased document would not contain any new information, and we treat them as *non-novel* instances. Table 4 shows an example of our interpretation of non-novelty in the dataset.

Table 4
Sample text from Webis-CPC-11 to simulate the high-level semantic paraphrasing in the dataset.

Original Text (Source Document)	**Paraphrase Text (Target Document: Non-Novel)**
The emigrants who sailed with Gilbert were better fitted for a crusade than a colony, and, disappointed at not at once finding mines of gold and silver, many deserted; and soon there were not enough sailors to man all the four ships. Accordingly, the Swallow was sent back to England with the sick; and with the remainder of the fleet, well supplied at St. John's with fish and other necessaries, Gilbert (August 20) sailed south as far as forty-four degrees north latitude. Off Sable Island, a storm assailed them, and the largest of the vessels, called the Delight, carrying most of the provisions, was driven on a rock and went to pieces.	The people who left their countries and sailed with Gilbert were more suited for fighting the crusades than for leading a settled life in the colonies. They were bitterly disappointed as it was not the America that they had expected. Since they did not immediately find gold and silver mines, many deserted. At one stage, there were not even enough men to help sail the four ships. So the Swallow was sent back to England carrying the sick. The other fleet was supplied with fish and the other necessities from St. John. On August 20, Gilbert had sailed as far as forty-four degrees to the north latitude. His ship known as the Delight, which bore all the required supplies, was attacked by a violent storm near Sable Island. The storm had driven it into a rock shattering it into pieces.

4.4.2 P4PIN Plagiarism Corpus. We use the P4PIN corpus (Sánchez-Vega 2016), a corpus especially built for evaluating the identification of paraphrase plagiarism. This corpus is an extension of the P4P corpus (Barrón-Cedeño et al. 2013), which contains pairs of text fragments where one fragment represents the original source text, and the other repre-

[3] https://www.uni-weimar.de/en/media/chairs/computer-science-department/webis/data/corpus.

Table 5
Sample from P4PIN to show plagiarism (non-novel) instance.

Original Text (Source Document)	Plagiarized Text (Target Document: Non-Novel)
I pored through these pages, and as I perused the lyrics of The Unknown Eros that I had never read before, I appeared to have found out something wonderful: there before me was an entire shining and calming extract of verses that were like a new universe to me.	I dipped into these pages, and as I read for the first time some of the odes of The Unknown Eros, I seemed to have made a great discovery: here was a whole glittering and peaceful tract of poetry, which was like a new world to me.

Table 6
Sample from Wikipedia Rewrite Dataset to show a plagiarism (non-novel) instance.

Original Text (Source Document)	Plagiarized Text (Target Document: Non-Novel)
PageRank is a link analysis algorithm used by the Google Internet search engine that assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set.	The PageRank algorithm is used to designate every aspect of a set of hyperlinked documents with a numerical weighting. The Google search engine uses it to estimate the relative importance of a web page according to this weighting.

sents a paraphrased version of the original. In addition, the P4PIN corpus includes *not paraphrase* plagiarism cases, that is, negative examples formed by pairs of unrelated text samples with likely thematic or stylistic similarity. The P4PIN dataset consists of 3,354 instances, 847 positives, and 2,507 negatives. We are interested in detecting plagiarism cases and also seeing the novelty scores for each category of instances predicted by our model. Table 5 represents a plagiarism (non-novel) example from P4PIN.

4.4.3 Wikipedia Rewrite Corpus. The dataset (Clough and Stevenson 2011) contains 100 pairs of short texts (193 words on average). For each of 5 questions about topics of computer science (e.g., "What is dynamic programming?"), a reference answer (source text, hereafter) has been manually created by copying portions of text from a relevant Wikipedia article. According to the degree of the rewrite, the dataset is 4-way classified as *cut & paste* (38 texts; a simple copy of text portions from the Wikipedia article), *light revision* (19; synonym substitutions and changes of grammatical structure allowed), *heavy revision* (19; rephrasing of Wikipedia excerpts using different words and structure), and *no plagiarism* (19; answer written independently from the Wikipedia article). We test or model on this corpus to examine the novelty scores predicted by our proposed approach for each category of answers. Please note that the information content for each of these answer categories is more or less the same as they cater to the same question. A sample from the dataset is shown in Table 6. For easier comprehension and fairer comparison, we accumulate some relevant dataset statistics in Table 7.

Table 7
Statistics of all the datasets. L →Average length of documents (sentences), Size→Size of the dataset in terms of number of documents. Emphasis is on detecting semantic-level non-novelty, which is supposedly more challenging than detecting novel texts.

Dataset	Objective	Size	L	Categories	Experimental Consideration
TAP-DLND 1.0	Novelty classification	6,109	∼15	Novel, non-novel	Each target document pitched against three source document.
TAP-DLND 2.0	Novelty scoring	8,271	∼16	Scores in the range 0 to 100	Each target document pitched against three source document.
Webis-CPC	Paraphrase detection	7,859	∼14	Paraphrase, non-paraphrase	Paraphrase instances as simulation of semantic-level non-novelty.
APWSJ	Novel document retrieval	11,896	∼28	Novel, partially redundant, absolutely redundant	Due to imbalance, partially redundant and absolutely redundant instances are together taken as non-novel.
P4PIN	Paraphrase plagiarism	3,354	∼3	Positive plagiarism, negative plagiarism	Plagiarism as a case for non-novelty.
Wikipedia Rewrite	Plagiarism	100	∼11	Cut-paste, light revision, heavy revision, no plagiarism	All categories simulate a kind of non-novelty at varying levels (lexical, semantic) of revision.

5. Evaluation

In this section, we evaluate the performance of our proposed approach, comparing it with baselines and also with our earlier approaches. We further show how our model performs in allied tasks like paraphrase detection, plagiarism detection, and identifying rewrites.

5.1 Baselines and Ablation Study

We carefully choose our baselines so that those also help in our ablation study. Baseline 1 emphasizes the role of textual entailment (i.e., what happens if we do not use the entailment principle in our model). With the Baseline 2 system, we investigate what happens if we do not include the *relevance detection module* in our architecture. Baseline 3 is similar to our earlier forays (Section 2.2) in the sense that we examine what happens if we do not assimilate information from multiple relevant premises and just fixate our attention to one single most relevant source premise. So, in essence, our Baseline systems 1, 2, 3 also signify our ablations on the proposed approach.

5.1.1 Baseline 1: Joint Encoding of Source and Target Documents. With this baseline, we want to see the importance of TE for our task of textual novelty detection. We use the Transformer variant of the Universal Sentence Encoder (Cer et al. 2018) to encode sentences in the documents to fixed-sized sentence embeddings (512 dimensions) and then stack them up to form the document embedding. We pass the source and target document representations to an MLP for corresponding feature extraction and final classification via softmax.

5.1.2 Baseline 2: Importance of Relevance Detection. With this baseline, we investigate the significance of relevance detection as a prior task to novelty detection. We turn off the relevance detection module and use the individual entailment decisions from the pre-trained ESIM model to arrive at the document-level aggregated decision.

5.1.3 Baseline 3: Single Premise. We keep all other parameters of our proposed model intact, but instead of having multiple premises, we take only the closest (top) premise (from the source sentences) for each target sentence. This way, we want to establish the importance of aggregating multiple premise entailment decisions for document-level novelty detection.

5.1.4 Baseline 4: Using BERT with MLP. We want to see how the recent state-of-the-art pre-trained large language models perform on our task. Essentially we use a BERT-base model (bert-base-uncased) with 12-layers, 12-attention-heads, and an embedding size of 768, for a total of 110M parameters and fine-tune on the novelty datasets in consideration. We feed the concatenation of source and target separated by [SEP] token into a pre-trained BERT (Bidirectional Encoder Representation from Transformers) (Devlin et al. 2019) model, then take the pooled output from the [CLS] token of the encoder and pass the representation so obtained to an MLP followed by classification via softmax. We take the implementation available in the HuggingFace library.[4] The original BERT model is pre-trained on the Toronto Book Corpus and Wikipedia. We keep the following hyperparameters during the task-specific (novelty detection) fine-tuning step: Learning rate: 2e-5, Num_train_epochs: 10, drop-out-rate: 0.1.

5.1.5 Baseline 5: Using a Simple Passage-level Aggregation Strategy. We follow a simple passage-level aggregation strategy as in Wang et al. (2018). We concatenate the selected source premises (top f) after the selection module to form the **union passage** of the premises (i.e., we do not scale with the relevance weights as in the original model) and then proceed next as per our proposed approach.

5.2 Comparing Systems

We compare with our earlier works on the same datasets, keeping all experimental configurations the same. A brief description of the prior work is in Section 2.2. Kindly refer to the papers for a detailed overview of the techniques.

5.2.1 Comparing System-1. With our first exploration on document-level novelty detection (Ghosal et al. 2018b), we use several features ranging from *lexical similarity, semantic similarity, divergence, keywords/NEs overlap, new word count*, and so on. The best-performing classifier was RF (Ho 1995). The idea was to exploit similarity and divergence-based handcrafted features for the problem. For more details on this comparing system, kindly refer to Section 2.2.1. This is the paper where we introduced the TAP-DLND 1.0 dataset for document-level novelty detection.

[4] `https://huggingface.co/transformers/model_doc/bert.html#bertmodel`.

5.2.2 Comparing System-2. With our next exploration, we introduce the concept of a RDV as a fused representation of the source and target documents (Ghosal et al. 2018a). We use a CNN to extract useful features for classifying the target document into novelty classes. For more details on this comparing system, kindly refer to Section 2.2.2.

5.2.3 Comparing System-3. To determine the amount of new information (novelty score) in a document, we generate a Source-Encapsulated Target Document Vector (SETDV) and train a CNN to predict the novelty score of the document (Ghosal et al. 2019). The value of the novelty score of a document ranges between 0 and 100 on the basis of new information content as annotated by our annotators (see Section 4.3). The architecture is quite similar to our RDV-CNN (Ghosal et al. 2018a), except that here, instead of classification, we are predicting the novelty score of the target document. The motivation here is that it is not always straightforward to ascertain what amount of newness makes a document appear novel to a reader. It is subjective and depends on the novelty appetite of the reader (Zhao and Lee 2016). Hence, we attempted to quantify *newness* for documents. The SETDV-CNN architecture also manifests the two-stage theory of human recall (Tulving and Kroll 1995) (search and retrieval, recognition) to select the most probable premise documents for a given target document.

5.2.4 Comparing System-4. With this work, we went on to explore the role of textual alignment (via decomposable attention mechanism) between target and source documents to produce a joint representation (Ghosal et al. 2021). We use a feed-forward network to extract features and classify the target document on the basis of new information content. For more details on this comparing system, kindly refer to Section 2.2.3.

5.3 BERT-NLI Variant of the Proposed Architecture

Because the contextual language models supposedly capture semantics better than the static language models, we experiment with a nearby variant in our proposed architecture. We make use of the BERT-based NLI model (Gao, Colombo, and Wang 2021) to examine the performance of BERT as the underlying language model in place of GloVe. This particular model is an instance of a NLI model, generated by fine-tuning Transformers on the SNLI and MultiNLI datasets (similar to ESIM). We use the same BERT-base variant as we do in Baseline 4. The rest of the architecture is similar to our proposed approach. We use the same BERT-based NLI model in the relevance module (to derive the relevance scores) and in the novelty detection module (for the final classification). We use the same configuration as Gao, Colombo, and Wang (2021) for fine-tuning the BERT-base on the NLI datasets.[5]

5.4 Hyperparameter Details

Our current architecture uses the ESIM stack as the entailment model pre-trained on SNLI and MultiNLI for both the relevance and novelty detection modules. Binary Cross Entropy is the loss function, and the default dropout is 0.5. We train for 10 epochs with Adam optimizer and keep the learning rate as 0.0004. The final feed-forward network has ReLU activation with a dropout of 0.2. The input size for the Bi-LSTM context encoder is 300 dimensions. We use the GloVe 800B embeddings for the input tokens. For

[5] https://github.com/yg211/bert_nli.

Table 8
Results on TAP-DLND 1.0. P→ Precision, R→ Recall, A→ Accuracy, R→ Recall, N→ Novel, NN→ Non-Novel, 10-fold cross-validation output, PLA→Passage-level Aggregation, as in Wang et al. (2018).

Evaluation System	**P(N)**	**R(N)**	F_1**(N)**	**P(NN)**	**R(NN)**	F_1**(NN)**	**A**
Baseline 1 (Joint Enc. Source+Target)	0.61	0.77	0.67	0.53	0.57	0.55	68.1%
Baseline 2 (w/o Relevance Detection)	0.84	0.57	0.67	0.71	0.86	0.77	76.4%
Baseline 3 (Single Premise)	0.82	0.70	0.76	0.77	0.84	0.80	80.3%
Baseline 4 (BERT+MLP)	0.84	0.87	0.85	0.88	0.89	0.88	87.0%
Baseline 5 (PLA)	0.89	0.68	0.77	0.73	0.91	0.81	79.7%
Comparing System 1 (Feature-based)	0.77	0.82	0.79	0.80	0.76	0.78	79.3%
Comparing System 2 (RDV-CNN)	0.86	**0.87**	**0.86**	0.84	0.83	0.83	84.5%
Comparing System 4 (Dec-Attn)	0.85	0.85	0.85	0.89	**0.89**	**0.89**	**87.4%**
Proposed Approach	**0.94**	0.77	0.85	0.80	**0.95**	0.87	87.2%
Proposed Approach (BERT-NLI)	**0.86**	0.87	0.86	0.88	**0.89**	0.89	**87.4%**

all uses of ESIM in our architecture, we initialize with the same pre-trained entailment model weights available with AllenNLP (Gardner et al. 2018).

5.5 Results

We discuss the results of our current approach in this section. We use TAP-DLND 1.0 and APWSJ datasets for our novelty classification experiments and the proposed TAP-DLND 2.0 dataset for quantifying new information experiments. We also report our experimental results on the Webis-CPC dataset, where we assume paraphrases to be simulating semantic-level non-novelty. We also show use cases of our approach for semantic-level plagiarism detection (another form of non-novelty in academia) with P4PIN and Wikipedia Rewrite datasets.

5.5.1 Evaluation Metrics. We keep the usual classification metrics for the novelty classification task: Precision, Recall, F_1 score, and Accuracy. For the APWSJ dataset, instead of accuracy, we report the Mistake (100-Accuracy) to compare with the earlier works. For the novelty scoring experiments on TAP-DLND 2.0, we evaluate our baselines and proposed model against the ground-truth scores using Pearson correlation coefficient, mean absolute error (the lower, the better), root mean squared error (the lower, the better), and the cosine similarity between the actual scores and the predicted scores.

5.5.2 On TAP-DLND 1.0 Dataset. Table 8 shows our results on TAP-DLND 1.0 dataset for the novelty classification task. As discussed, in Section 3.2, here we keep $f = 10$, that is, the topmost ten relevant source sentences (based on α_{kf} scores) as the relevant premises for each target sentence t_k in the target document. We can see that our current approach performs comparably with our preceding approach (Comparing System 4). With a high recall for non-novel class, we can say that our approach has an affinity to discover document-level *non-novelty*, which is comparatively more challenging at the semantic level. The results in Table 9 are from 10-fold cross-validation experiments.

5.5.3 On APWSJ Dataset. The APWSJ dataset is more challenging than TAP-DLND 1.0 because of the sheer number of preceding documents one has to process for deciding the

Table 9
Results for redundant class on APWSJ. *Mistake* →100-Accuracy. Except for Zhang, Callan, and Minka (2002b), all other results correspond to a 10-fold cross-validation output.

Measure	**Recall**	**Precision**	**Mistake**
Baseline 1 (Joint Enc. Source+Target)	0.66	0.75	28.8%
Baseline 2 (w/o Relevance Detection)	0.76	0.85	18.8%
Baseline 3 (Single Premise)	0.85	0.86	13.4%
Baseline 4 (BERT+MLP)	0.87	0.90	8.2%
Baseline 5 (PLA: (Wang et al. 2018))	0.78	0.88	18.2%
Comparing System (Zhang, Callan, and Minka 2002b)	0.56	0.67	27.4%
Comparing System 2 (RDV-CNN)	0.58	0.76	22.9%
Comparing System 4 (Dec-Attn)	0.86	0.92	7.8%
Proposed Approach	**0.91**	**0.95**	**5.9%**
Proposed Approach (BERT-NLI)	0.90	0.93	6.2%

state of the novelty of the current one. The first document in the chronologically ordered set of documents for a given topic is always *novel* as it starts the story. The novelty of all other documents is judged based on the chronologically preceding ones. Thus for the final document in a given topic (see Section 4.2 for the TREC topics), the network needs to process all the preceding documents in that topic. Although APWSJ was developed from an information retrieval perspective, we take a classification perspective (i.e., to classify the current document into *novel* or *non-novel categories* based on its chronological priors) for our experiments. Table 9 reports our result and compares it with earlier systems. Kindly note that we take the same experimental condition as the original paper (Zhang, Callan, and Minka 2002b) and consider *partially-redundant* documents into the *redundant* class. Our current approach performs much better than the earlier reported results with $f = 10$, thereby signifying the importance of multi-premise entailment for the task at hand. We report our results on the redundant class as in earlier systems. Finding semantic-level non-novelty for documents is much more challenging than identifying whether a document has enough new things to say to classify it as *novel*.

5.5.4 On TAP-DLND 2.0 Dataset. On our newly created dataset for predicting novelty scores, instead of classification we try to squash the output to a numerical score. We use the same architecture in Figure 1 but use sigmoid activation at the last layer to restrict the score within the range of 100. Table 10 shows our performance. This experiment is particularly important to quantify the amount of *newness* in the target document with respect to the source documents. Kindly note we allow a +5 and −5 range with respect to the human-annotated score for our predicted scores. We see that our current approach performs comparably with the earlier reported results.

5.5.5 Ablation Studies. As we mentioned, our baselines serve as means of ablation studies. Baseline 1 is the simplest one where we simply let the network discover useful features from the universal representations of the *source-target* pairs. We do not apply any sophisticated approach, and it performs the worst. Baseline 1 establishes the importance of our TE pipeline in the task. In Baseline 2, we do not consider the *relevance* detection module and hence do not include the relevance weights in the architecture. Baseline 2 performs much better than Baseline 1 (relative improvement of 8.3% in the TAP-DLND

Table 10
Performance of the proposed approach against the baselines and comparing systems TAP-DLND 2.0. PC→ Pearson Correlation Coefficient, MAE→ Mean Absolute Error, RMSE→ Root Mean-Squared Error, Cosine→ Cosine similarity between predicted and actual score vectors. Comparing System 2 and 3 are thematically the same.

Evaluation System	**PC**	**MAE**	**RMSE**	**Cosine**
Baseline 1 (Joint Enc. Source+Target)	0.69	36.11	49.92	0.87
Baseline 2 (w/o Relevance Detection)	0.81	15.34	23.83	0.91
Baseline 3 (Single Premise)	0.84	12.40	20.14	0.93
Baseline 4 (BERT+MLP)	0.82	15.44	22.21	0.93
Baseline 5 (PLA: (Wang et al. 2018))	0.84	11.78	18.06	0.92
Comparing System 3 (SETDV-CNN)	0.88	**10.29**	**16.54**	**0.95**
Comparing System 4 (Dec-Attn)	0.61	31.07	26.51	0.81
Proposed Approach	**0.88**	10.92	17.73	0.94
Proposed Approach (BERT-NLI)	**0.88**	10.42	17.32	0.94

1.0 dataset and minimizing mistakes to the extent of 10% for APWSJ). For Baseline 3, we take only the single most relevant premise (having the highest relevance score) instead of multiple premises. It improves over Baseline 2 by a margin of 3.9% for TAP-DLND 1.0 and 5.2% for APWSJ. We observe almost similar behavior for novelty-scoring in TAP-DLND 2.0. However, with our proposed approach, we attain significant performance gain over our ablation baselines, as is evident in Tables 8, 9, and 10. Thus our analysis indicates the importance of having *relevance scores* in a *multi-premise* scenario for the task at hand.

5.6 Results on Related Tasks

To evaluate the efficacy of our approach, we went ahead to test our model on certain related tasks to textual novelty (Section 4.4).

5.6.1 Paraphrase Detection. As already mentioned, paraphrase detection is one such task that simulates the notion of non-novelty at the semantic level. Detecting semantic-level redundancies is not straightforward. We are interested in identifying those documents that are lexically distant from the source yet convey the same meaning (thus semantically non-novel). For our purpose, we experiment with the Webis-CPC-11 corpus, which consists of paraphrases from high-level literary texts (see Table 4, for example, simulating non-novelty). We report our results on the paraphrase class as the non-paraphrase instances in this dataset do not conform to novel documents. We perform comparably with our earlier results (Table 11). This is particularly encouraging because detecting semantic-level non-novelty is challenging, and the quality of texts in this dataset is richer than more straightforward newspaper texts (Table 4).

5.6.2 Plagiarism Detection. We envisage plagiarism as one form of semantic-level non-novelty. We discuss our performance on plagiarism detection below.

Table 11
Results for paraphrase class on Webis-CPC, 10-fold cross-validation output.

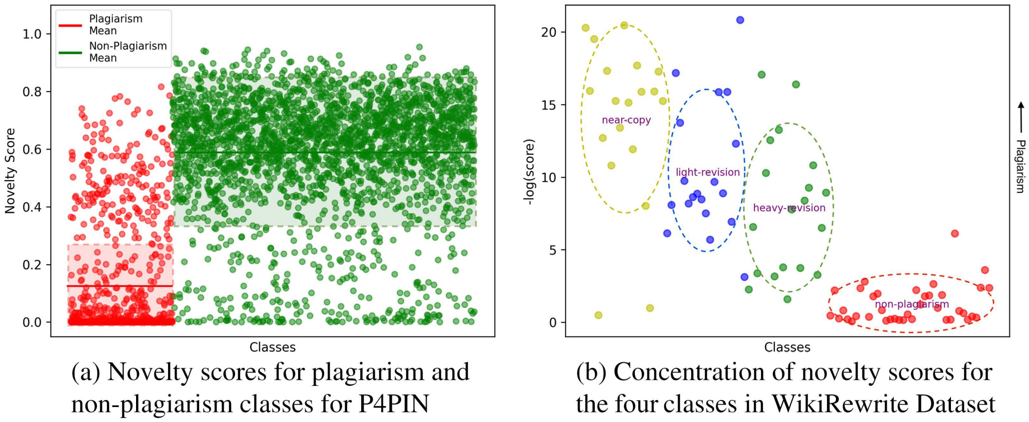
Evaluation System	P	R	F_1	A
Baseline 1 (Joint Enc. Source+Target)	0.58	0.69	0.63	58.0%
Baseline 2 (w/o Relevance Detection)	0.73	0.92	0.81	77.6%
Baseline 3 (Single Premise)	0.74	0.92	0.82	78.2%
Baseline 4 (BERT+MLP)	0.85	0.75	0.79	78.2%
Baseline 5 (PLA: (Wang et al. 2018))	0.93	0.57	0.71	77.8%
Comparing System 2 (RDV-CNN)	0.75	0.84	0.80	78.0%
Comparing System 4 (Dec-Attn)	0.72	0.88	0.79	76.4%
Proposed Approach	0.76	**0.90**	**0.82**	**78.9%**
Proposed Approach (BERT-NLI)	**0.85**	0.88	**0.86**	**82.1%**

P4PIN Dataset

Semantic-level plagiarism is another task that closely simulates non-novelty. The P4PIN dataset is not large (only 847 plagiarism instances) and is not suitable for a deep learning experiment setup. We adapt a *transfer learning* scheme and train our model on TAP-DLND 1.0 (novelty detection task), and test if our model can identify the plagiarism cases in P4PIN. We are not interested in the *non-plagiarism* instances as those do not conform to our idea of *novelty*. Non-plagiarism instances in P4PIN exhibit thematic and stylistic similarity to the content of the original text. We correctly classify 832 out of 847 plagiarized instances, yielding a sensitivity of 0.98 toward identifying semantic-level plagiarism. Figure 5a shows the predicted novelty scores for the documents in P4PIN (trained on TAP-DLND 2.0). We can clearly see that the concentration of novelty scores for the plagiarism class is at the bottom, indicating low novelty, while that for the non-plagiarism class is at the upper half, signifying higher novelty scores.

Wikipedia Rewrite

We also check how our model can identify the various degree of rewrites (plagiarism) with the Wikipedia Rewrite Dataset. Here again, we train on TAP-DLND 2.0. We take the negative log of the predicted scores (the higher the result, the less is the novelty score) and plot along the y-axis in Figure 5b. According to our definition, all the four classes of documents (*near-copy, light-revision, heavy-revision, non-plagiarism*) are not novel. But the degree of non-novelty should be higher for *near copy*, followed by *light revision*, and then *heavy revision*. *Near Copy* simulates a case of lexical-level plagiarism whereas *light revision* and *heavy revision* could be thought of as plagiarism at the semantic-level. The novelty scores predicted by our model display the novelty score concentration in clusters for each category. If there is no plagiarism, the novelty score is comparatively higher (non-plagiarism instances are at the bottom signifying higher novelty scores). All these performances of our approach on prediction of the non-novel instances indicates that finding multiple sources and assimilating the corresponding information to arrive at the judgement for novelty/non-novelty is essential.



(a) Novelty scores for plagiarism and non-plagiarism classes for P4PIN

(b) Concentration of novelty scores for the four classes in WikiRewrite Dataset

Figure 5
Predicted novelty scores for documents in P4PIN and WikiRewrite by our model trained on TAP-DLND 1.0.

5.7 On Using Contextual Language Models

It is quite evident from our experiments that the recent pre-trained large contextual language models (BERT in particular) with a simple architecture performs well with the concerned task (Baseline 4). The BERT-NLI version of the GloVe-based ESIM stack modeled as per our proposed approach performs comparably, sometimes even better. Especially the BERT-NLI version of our proposed approach performs better in identifying semantic-level redundancies (non-novelty, paraphrases). We assume that it would be an interesting direction to use the very large billion parameter language models (like T5 [Raffel et al. 2020], GPT3 [Brown et al. 2020], Megatron-Turing Natural Language Generation,[6] etc.) to automatically learn the notion of *newness* from the source-target itself.

The passage-level aggregation baseline (Wang et al. 2018) performed comparatively better than the other baselines; however, the proposed approach edged it. This is probably due to scaling the selected premise representations by their corresponding relevance scores.

5.8 Analysis

The actual documents in all of our datasets are long and would not fit within the scope of this article. Hence, we take the same example in Section 1 (Example 2) to analyze the performance of our approach.

Figure 6 depicts the heatmap of the attention scores between the target and source document sentences. We can clearly see that for target sentence t_1 the most relevant source sentences predicted by our model are s_1, s_2, s_3. While we read t_1 (*Facebook*

[6] https://tinyurl.com/megatron-nvidia.

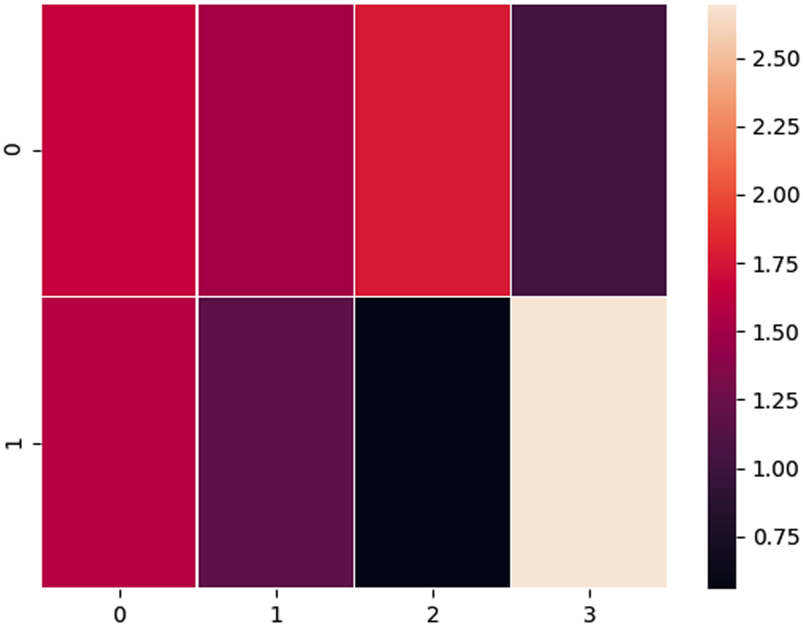


Figure 6
Heatmap depicting the attention scores between the source and target document (Example 2 in Section 1). t_1, t_2 are the target document sentences (vertical axes), and s_1, s_2, s_3, s_4 are source document sentences (horizontal axes). The brighter the shade, the more is the alignment, signifying an affinity toward non-novelty.

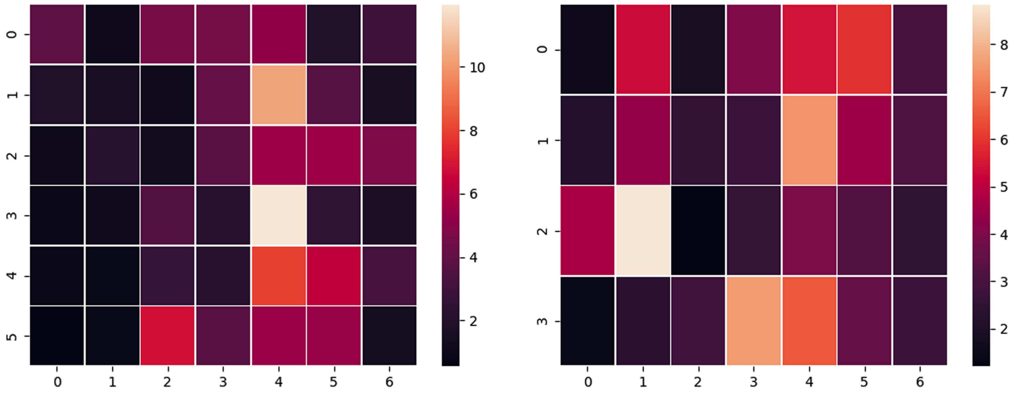
was launched in Cambridge) against the source document, we can understand that t_1 is offering no new information. But in order to do that we need to do a multi-hop reasoning over s_1 (*Facebook*) $\rightarrow$ s_2 (*created in Harvard*) $\rightarrow$ s_3 (*Harvard is in Cambridge*). The other information in s_4 (*Zuckerberg lives in California*) does not contribute to ascertaining t_1 and hence is a distracting information. Our model pays low attention to s_4.

Similarly, when we consider the next target sentence t_2 (*The founder resides in California*), we understand that s_4 (*Zuckerberg lives in California*), s_2 (*Zuckerberg created Facebook*), and s_1 (*Facebook*) are the source sentences, which ascertains that t_2 does not have any new information. s_3 (*Harvard is in Cambridge*) finds no relevance to the sentence in concern. Hence our model assigns lowest attention score to s_3 for t_2, signifying that s_3 is a distracting premise.

Finally, our model predicts that the target document in concern is *non-novel* with respect to the source document. The predicted *novelty-score* was 20.59 on a scale of 100. Let us now take a more complicated example.

Source Document 1 (S_1): Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people who fall sick with COVID-19 will experience mild to moderate symptoms and recover without special treatment.

Source Document 2 (S_2): The virus that causes COVID-19 is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales. These droplets are too heavy to hang in the air and quickly fall on floors or surfaces. You can be infected by breathing in the virus if you are within close proximity of someone who has COVID-19 or by touching a contaminated surface and then your eyes, nose, or mouth.



(a) Heatmap for Target T_1 against S_1, S_2, S_3 (b) Heatmap for Target T_2 against S_1, S_2, S_3

Figure 7
Heatmap depicting the attention scores between the source (S_1, S_2, S_3) and target document (T_1, T_2). The brighter the shade, the more is the alignment, signifying an affinity toward non-novelty.

Source Document 3 (S_3): You can reduce your chances of being infected or spreading COVID-19 by regularly and thoroughly cleaning your hands with an alcohol-based hand rub or washing them with soap and water. Washing your hands with soap and water or using alcohol-based hand rub kills viruses that may be on your hands.

Target T_1 (Non-Novel): Coronavirus is a respiratory illness, meaning it is mainly spread through virus-laden droplets from coughs and sneezes. The government's advice on Coronavirus asks the public to wash their hands more often and avoid touching their eyes, nose, and mouth. Hands touch many surfaces and can pick up viruses. Once contaminated, hands can transfer the virus to your eyes, nose, or mouth. From there, the virus can enter your body and infect you. You can also catch it directly from the coughs or sneezes of an infected person.

Target T_2: COVID-19 symptoms are usually mild and begin gradually. Some people become infected but don't develop any symptoms and feel unwell. Most people (about 80%) recover from the disease without needing special treatment. Older people, and those with underlying medical problems like high blood pressure, heart problems or diabetes, are more likely to develop serious illnesses.

The heatmap for the above examples after prediction is shown in Figure 7. Keeping the source documents (S_1, S_2, S_3) the same, we analyze our model's prediction against the two Target Documents (T_1 and T_2). The source document sentences are along the horizontal axes, while the target document sentences are along the vertical axes. After reading T_1 and T_2 against S_1, S_2, S_3 we can understand that T_1 is offering very little new information, however T_2 has some amount of new information (*Older people are more susceptible to the disease*). Our model predicts 22.73 and 40.30 as novelty scores for T_1 and T_2, respectively, which is somewhat intuitive. Intuitively, both the target documents T_1 and T_2 appears *non-novel* with respect to the source documents S_1, S_2, and S_3.

The third sentence in T_2 (*Most people (about 80%) recover from the disease without needing special treatment*) highly attends the second sentence in S_1 (*Most people who fall sick with COVID-19 will experience mild to moderate symptoms and recover without special treatment*). Similarly, the third sentence in S_2 pays greater attention to the fourth sentence in T_1, signifying that the target sentence has less/no new information with respect to the source candidates.

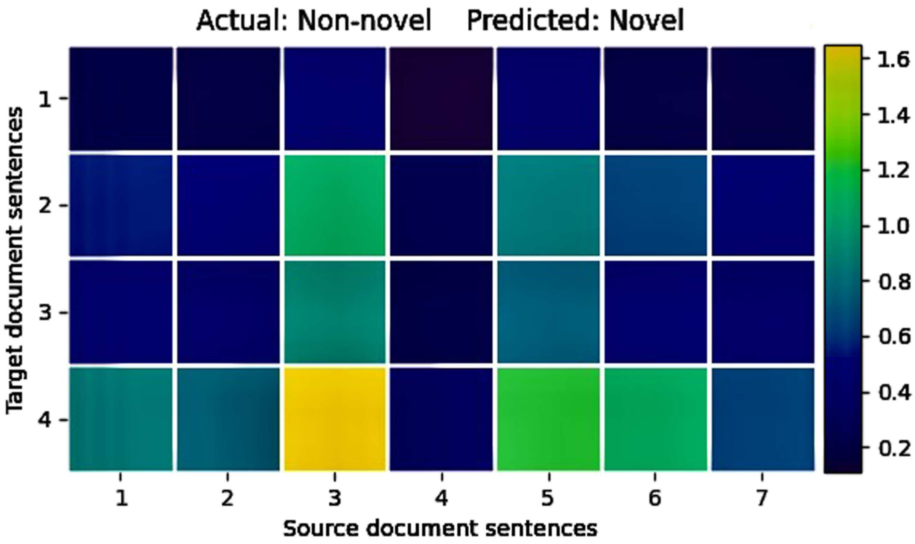


Figure 8
Heatmap of attention values from the decomposable attention-based model for novelty detection (Comparing System 4) for the Target T_2 against the Source documents S_1, S_2, S_3. Due to low attention values, the model predicts the document pair as 'Novel', which is not correct.

We can see via the above heatmap figures how multiple premises in the source documents are attending the target sentences, which is correctly captured by our approach, hence establishing our hypothesis. We also experiment with our earlier best-performing model, Comparing System 4: Decomposable attention-based novelty detection. However, the decomposable attention-based model predicts the class incorrectly, as we can see in Figure 8. The model assigns low attention values between the source-target pair sentences, hence predicting the target document as *novel*. However, our current approach correctly predicts the class label of the target document.

5.9 Error Analysis

We have identified a few causes of errors committed by our approach.

- **Long Documents:** The misclassified instances in the datasets (APWSJ, TAP-DLND 1.0) are too long. Also, the corresponding source documents have a good amount of information. Although our architecture works at sentence-level and then composes at the document level, finding the relevant premises from large documents is challenging.
- **Non-coherence of Premises:** Another challenge is to aggregate the premises as the premises are not in a coherent order after selection in the Selection Module.
- **Named Entities:** Let us consider a misclassified instance (see the heatmap in Figure 9) with respect to the COVID-19 source documents in the earlier example.

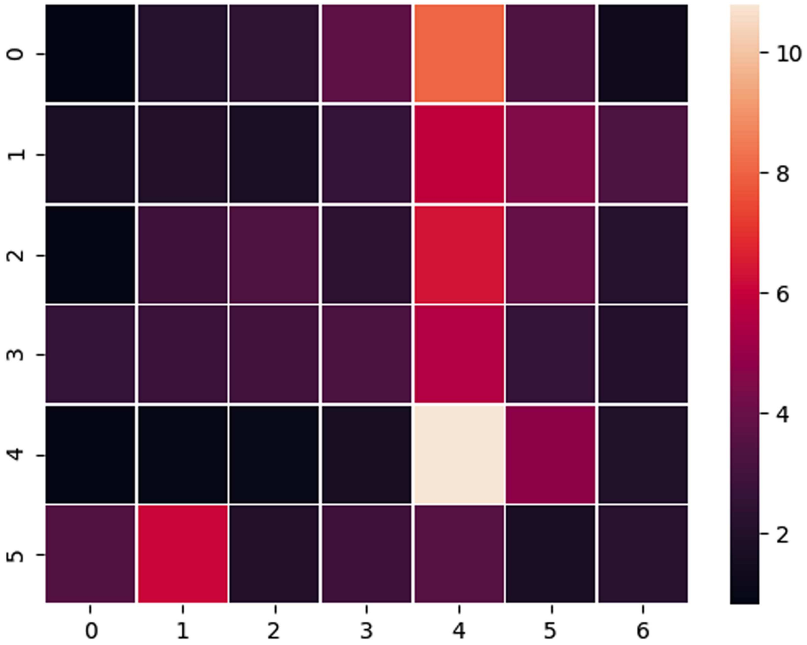


Figure 9
Heatmap of the misclassification instance.

Target T_3 (Novel): The world has seen the emergence of a Novel Corona Virus on 31 December 2019, officially referred to as COVID-19. The virus was first isolated from persons with pneumonia in Wuhan city, China. The virus can cause a range of symptoms, ranging from mild illness to pneumonia. Symptoms of the disease are fever, cough, sore throat, and headaches. In severe cases, difficulty in breathing and deaths can occur. There is no specific treatment for people who are sick with Coronavirus and no vaccine to prevent the disease.

We could clearly understand that T_3 has new information with respect to the source documents. But due to higher correspondence in NEs and certain content words (e.g., virus) between source-target pairs, our classifier may have got confused and predicted T_3 as non-novel. Kindly note that our documents in the actual dataset are much longer than the examples we demonstrate, adding more complexity to the task.

6. Summary, Conclusion, and Future Work

Textual Novelty Detection has an array of use-cases starting from search and retrieval on the Web, NLP tasks like plagiarism detection, paraphrase detection, summarization, modeling interestingness, fake news detection, and so forth. However, less attention is paid to the document-level variant of the problem in comparison to sentence-level novelty detection. In this work, we present a comprehensive account of our experiments so far on *document-level novelty detection*. We study existing literature on textual novelty detection as well as our earlier explorations on the topic. Here we assert that we would need to perform information assimilation from multiple premises to identify the novelty of a given text. Our current approach performs better than our earlier approaches. Also, we show that our method could be suitably applied to allied tasks like Plagiarism Detection and Paraphrase Detection. We point out some limitations of our approach, which we aim to explore next.

In the future, we would aim to explore novelty detection in scientific texts, which would be much more challenging than newspaper texts. We would also like to investigate how we could address situations when the number of source documents increases exponentially. An interesting direction to probe next would be to understand the subjectivity associated with the task across multiple human raters to understand better how *newness* is perceived by humans under different conditions. This would also help understand and probably eliminate any human biases toward the novelty labeling that may have accidentally crept in. We make our data and codes are available at `https://github.com/Tirthankar-Ghosal/multipremise-novelty-detection`.

Acknowledgments

This work sums up one chapter of the dissertation of the first author. The current work draws inspiration from our earlier works published in LREC 2018, COLING 2018, IJCNN 2019, and NLE 2020. We acknowledge the contributions and thank the several anonymous reviewers for their suggestions to take up this critical challenge and improve our investigations. We thank our annotators, Ms. Amitra Salam and Ms. Swati Tiwari, for their commending efforts to develop the dataset. We also thank the Visvesvaraya Ph.D. Scheme of Digital India Corporation under the Ministry of Electronics and Information Technology, Government of India, for providing Ph.D. fellowship to the first author and faculty award to the fourth author to do our investigations on Textual Novelty. Dr. Asif Ekbal acknowledges the Visvesvaraya Young Faculty Research Fellowship (YFRF) Award, supported by the Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for this research.

References

Ahmad, Amin, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. ReQA: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019*, pages 137–146, `https://doi.org/10.18653/v1/D19-5819`

Allan, James, Victor Lavrenko, Daniella Malin, and Russell Swan. 2000. Detections, bounds, and timelines: Umass and TDT-3. In *Proceedings of Topic Detection and Tracking Workshop*, pages 167–174.

Allan, James, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45.

Allan, James, Courtney Wade, and Alvaro Bolivar. 2003a. Retrieval and novelty detection at the sentence level. In *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 314–321. `https://doi.org/10.1145/860435.860493`

Allan, James, Courtney Wade, and Alvaro Bolivar. 2003b. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 314–321, ACM.

Augenstein, Isabelle, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 4684–4696. `https://doi.org/10.18653/v1/D19-1475`

Bagga, Amit and Breck Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *Coreference and Its Applications*.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, pages 150–165.

Barrón-Cedeño, Alberto, Marta Vila, Maria Antònia Martí, and Paolo Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947. https://doi.org/10.1162/COLI_a_00153

Bentivogli, L., P. Clark, I. Dagan, H. T. Dang, and D. Giampiccolo 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge. In *In TAC 2011 Notebook Proceedings*, pges 1–16.

Bentivogli, L., P. Clark, I. Dagan, and D. Giampiccolo 2010. The Sixth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Text Analysis Conference (TAC 2010)*, pages 1–60.

Bernstein, Yaniv and Justin Zobel. 2005. Redundant documents and search effectiveness. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 736–743.

Bhatnagar, Vasudha, Ahmed Sultan Al-Hegami, and Naveen Kumar. 2006. Novelty as a measure of interestingness in knowledge discovery. *Constraints*, 9:18.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. https://doi.org/10.18653/v1/D15-1075

Brants, Thorsten, Francine Chen, and Ayman Farahat. 2003. A system for new event detection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 330–337.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33:1877–1901.

Burrows, Steven, Martin Potthast, and Benno Stein. 2013. Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):43.

Bysani, Praveen. 2010. Detecting novelty in the context of progressive summarization. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 13–18.

Carbonell, Jaime and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations*, pages 169–174. https://doi.org/10.18653/v1/d18-2029

Chandar, Praveen and Ben Carterette. 2013. Preference based evaluation measures for novelty and diversity. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 413–422. https://doi.org/10.1145/2484028.2484094

Chen, Qian, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668. https://doi.org/10.18653/v1/P17-1152

Chen, Tongfei, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 8772–8779. https://doi.org/10.18653/v1/2020.acl-main.774

Clarke, Charles L. A., Nick Craswell, Ian Soboroff, and Azin Ashkan. 2011. A comparative analysis of cascade measures for novelty and diversity, WSDM '11, pages 75–84. https://doi.org/10.1145/1935826.1935847

Clarke, Charles L. A., Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 659–666. https://doi.org/10.1145/1390334.1390446

Clough, Paul D. and Mark Stevenson. 2011. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45(1):5–24. https://doi.org/10.1007/s10579-009-9112-1

Collins-Thompson, Kevyn, Paul Ogilvie, Yi Zhang, and Jamie Callan. 2002. Information filtering, novelty detection, and named-page finding. In *TREC*, pages 1–12.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. 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*, pages 670–680.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190, Springer. https://doi.org/10.1007/11736790_9

Dagan, Ido, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Dasgupta, Tirthankar and Lipika Dey. 2016. Automatic scoring for innovativeness of textual ideas. In *Knowledge Extraction from Text, Papers from the 2016 AAAI Workshop*, pages 6–11.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423

Du, Jingfei, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418. https://doi.org/10.18653/v1/2021.naacl-main.426

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters, *Psychological Bulletin*, 76(5):378.

Franz, Martin, Abraham Ittycheriah, J. Scott McCarley, and Todd Ward. 2001. First story detection: Combining similarity and novelty based approaches. In *Topic Detection and Tracking Workshop Report*, pages 193–206.

Gabrilovich, Evgeniy, Susan Dumais, and Eric Horvitz. 2004. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th International Conference on World Wide Web*, pages 482–490.

Gamon, Michael. 2006. Graph-based text representation for novelty detection. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 17–24.

Gao, Yang, Nicolò Colombo, and Wei Wang. 2021. Adapting by pruning: A case study on BERT. *CoRR*, abs/2105.03343:66–78.

Gardner, Matt, 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. https://doi.org/10.18653/v1/W18-2501

Ghosal, Tirthankar, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, Srinivasa Satya Sameer Kumar Chivukula, and George Tsatsaronis. 2021. Is your document novel? Let attention guide you. An attention based model for document-level novelty detection. *Natural Language Engineering*, 27(4):427–454. https://doi.org/10.1017/S1351324920000194

Ghosal, Tirthankar, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, George Tsatsaronis, and Srinivasa Satya Sameer Kumar Chivukula. 2018a. Novelty goes deep. A deep neural solution to document level novelty detection. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 2802–2813.

Ghosal, Tirthankar, Amitra Salam, Swati Tiwary, Asif Ekbal, and Pushpak Bhattacharyya. 2018b. 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*, pages 3541–3547. `https://aclanthology.org/L18-1559`

Ghosal, Tirthankar, Abhishek Shukla, Asif Ekbal, and Pushpak Bhattacharyya. 2019. To comprehend the new: On measuring the freshness of a document. In *International Joint Conference on Neural Networks, IJCNN 2019*, pages 1–8. `https://doi.org/10.1109/IJCNN.2019.8851857`

Gipp, Bela, Norman Meuschke, and Corinna Breitinger. 2014. Citation-based plagiarism detection: Practicability on a large-scale scientific corpus. *Journal of the Association for Information Science and Technology*, 65(8):1527–1540.

Harman, Donna. 2002a. Overview of the TREC 2002 novelty track. In *Proceedings of The Eleventh Text REtrieval Conference, TREC 2002*, pages 1–20.

Harman, Donna. 2002b. Overview of the TREC 2002 novelty track. In *TREC*, pages 46–55.

Ho, Tin Kam. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282, IEEE.

Huang, Qiang, Jianhui Bu, Weijian Xie, Shengwen Yang, Weijia Wu, and Liping Liu. 2019. Multi-task sentence encoding model for semantic retrieval in question answering systems. In *International Joint Conference on Neural Networks, IJCNN 2019*, pages 1–8, IEEE. `https://doi.org/10.1109/IJCNN.2019.8852327`

Jaccard, Paul. 1901. Étude comparative de la distribution florale dans une portion des alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Karkali, Margarita, François Rousseau, Alexandros Ntoulas, and Michalis Vazirgiannis. 2013. Efficient online novelty detection in news streams. In *Web Information Systems Engineering - WISE 2013 - 14th International Conference, Proceedings, Part I*, pages 57–71. `https://doi.org/10.1007/978-3-642-41230-1_5`

Kim, Yoon. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Kwee, Agus T., Flora S. Tsai, and Wenyin Tang. 2009. Sentence-level novelty detection in English and Malay. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 40–51.

Lai, Alice, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises, Greg Kondrak and Taro Watanabe, editors. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Volume 1: Long Papers*, pages 100–109.

Li, Xiaoyan and W. Bruce Croft. 2005. Novelty detection based on sentence level patterns. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 744–751. `https://doi.org/10.1145/1099554.1099734`

Lin, Chin Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Mihalcea, Rada and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004*, pages 404–411, ACL.

Mou, Lili, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 130–136. `https://doi.org/10.18653/v1/P16-2022`

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Parikh, Ankur, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255. https://doi.org/10.18653/v1/D16-1244

Pavlick, Ellie and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Qin, Yumeng, Dominik Wurzer, Victor Lavrenko, and Cunchen Tang. 2016. Spotting rumors via novelty detection. *CoRR*, abs/1611.06322:1–12.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journals of Machine Learning Research*, 21:140:1–140:67.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. https://doi.org/10.18653/v1/D16-1264

Ru, Liyun, Le Zhao, Min Zhang, and Shaoping Ma. 2004. Improved Feature Selection and Redundance Computing - THUIR at TREC 2004 Novelty Track. *TREC*, volume 500-261, pages 1–14.

Saikh, Tanik, Tirthankar Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. 2017. Document level novelty detection: Textual entailment lends a helping hand. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 131–140.

Sánchez-Vega, José Fernando. 2016. *Identificación de plagio parafraseado incorporando estructura, sentido y estilo de los textos*. PhD thesis, Instituto Nacional de Astrofísica, Optica y Electrónica.

Schiffman, Barry and Kathleen R. McKeown. 2005. Context and learning in novelty detection. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 716–723.

Soboroff, Ian. 2004. Overview of the TREC 2004 novelty track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*.

Soboroff, Ian and Donna Harman. 2003. Overview of the TREC 2003 novelty track. In *TREC*, pages 38–53.

Soboroff, Ian and Donna Harman. 2005. Novelty detection: The TREC experience. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 105–112.

Stokes, Nicola and Joe Carthy. 2001. First story detection using a composite document representation. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8.

Tarnow, Eugen. 2015. First direct evidence of two stages in free recall. *RUDN Journal of Psychology and Pedagogics*, (4):15–26.

Trivedi, Harsh, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering tasks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 2948–2958. https://doi.org/10.18653/v1/n19-1302

Tsai, Flora S. and Kap Luk Chan. 2010. Redundancy and novelty mining in the business blogosphere. *The Learning Organization*, 17(6):490–499.

Tsai, Flora S., Wenyin Tang, and Kap Luk Chan. 2010. Evaluation of novelty metrics for sentence-level novelty mining. *Information Sciences*, 180(12):2359–2374.

Tsai, Flora S. and Yi Zhang. 2011. D2s: Document-to-sentence framework for novelty detection. *Knowledge and Information Systems*, 29(2):419–433. https://doi.org/10.1007/s10115-010-0372-2

Tulving, Endel and Neal Kroll. 1995. Novelty assessment in the brain and long-term memory encoding. *Psychonomic Bulletin & Review*, 2(3):387–390.

Verheij, Arnout, Allard Kleijn, Flavius Frasincar, and Frederik Hogenboom. 2012. A comparison study for novelty control mechanisms applied to Web news stories. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences*, volume 1, pages 431–436.

Wang, Shuohang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2018.

Evidence aggregation for answer re-ranking in open-domain question answering. In *6th International Conference on Learning Representations, ICLR 2018, Conference Track Proceedings*, pages 1–16, OpenReview.net.

Wayne, Charles L. 1997. Topic Detection and Tracking (TDT). In *Workshop held at the University of Maryland*, volume 27, page 28. Citeseer.

Williams, Adina, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Yang, Yiming, Tom Pierce, and Jaime Carbonell. 1998. 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*, pages 28–36.

Yang, Yiming, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 688–693. `https://doi.org/10.1145/775047.775150`

Yang, Yinfei, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020*, pages 87–94. `https://doi.org/10.18653/v1/2020.acl-demos.12`

Yang, Zhilin, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. `https://doi.org/10.18653/v1/d18-1259`

Zhang, Min, Ruihua Song, Chuan Lin, Shaoping Ma, Zhe Jiang, Yijiang Jin, Yiqun Liu, Le Zhao, and S. Ma. 2003. Expansion-based technologies in finding relevant and new information: THU TREC 2002: Novelty Track Experiments. *NIST Special Publication SP*, (251):586–590.

Zhang, Yi, Jamie Callan, and Thomas Minka. 2002a. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–88.

Zhang, Yi, James P. Callan, and Thomas P. Minka. 2002b. 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*, pages 81–88. `https://doi.org/10.1145/564376.564393`

Zhang, Yi and Flora S. Tsai. 2009. Combining named entities and tags for novel sentence detection. In *Proceedings of the WSDM09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 30–34, ACM.

Zhao, Pengfei and Dik Lun Lee. 2016. How much novelty is relevant?: It depends on your curiosity. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–324, ACM.