TAP-DLND 1.0 : A Corpus for Document Level Novelty Detection

Tirthankar Ghosal, Amitra Salam, Swati Tiwari, Asif Ekbal, Pushpak Bhattacharyya
Indian Institute of Technology Patna
Bihta, Bihar-801106, India
(tirthankar.pcs16, salam.recs, stlex.re, asif, pb)@iitp.ac.in

Abstract

Detecting novelty of an entire document is an Artificial Intelligence (AI) frontier problem. This has immense importance in widespread Natural Language Processing (NLP) applications ranging from extractive text document summarization to tracking development of news events to predicting impact of scholarly articles. Although a very relevant problem in the present context of exponential data duplication, we are unaware of any document level dataset that correctly addresses the evaluation of automatic novelty detection techniques in a classification framework. To bridge this relative gap, here in this work, we present a resource for benchmarking the techniques for *document level novelty detection*. We create the resource via *topic-specific crawling* of news documents across several domains in a periodic manner. We release the annotated corpus with necessary statistics and show its use with a developed system for the problem in concern.

Keywords: Document level novelty detection, Classification, Web crawling, Corpus

1. Introduction

Novelty detection implies finding elements that have not appeared before, or new, or original with respect to relevant references. The explosive growth of documents across the web has resulted in the accumulation of redundant ones, thereby consuming space as well as precious time of readers seeking new information. This necessitates finding means for discarding redundant document(s) and retaining ones containing novel information. The level of information duplication is not just limited to the lexical surface form of texts but has encroached the barriers of semantics and pragmatics too. Paraphrasing, semantic level plagiarism etc. are instances of such practices. Intelligent text reuse, synonym replacement and careful alignment may lead to a surface form which is very different from the originating source yet convey the same meaning. Present state-of-the-art text matching techniques are unable to process such redundancy. The quest of new information is an eternal human need and urges attention in this very age of exploratory data redundancy. One major objective of this work is to provide a benchmark setup for experiments to filter out superfluous information across the web. With this work we introduce a simplistic dataset to the research community to inculcate efficient methods to detect *document level novelty* or on the contrary document level redundancy. We create the resource by crawling news articles of different categories by various agencies and coin it as *TAP-DLND 1.0*[1] (after the initial names of the principal investigators *Tirthankar-Asif-Pushpak*) which also stands for *Explore Document Level Novelty Detection (DLND)*. In this work we view the problem of novelty detection as a two-class classification problem with the judgment that whether an incoming document bears sufficiently new information to be labeled as novel with respect to a set of source documents. The source document set could be seen as the memory of the reader which stores known information. We extract features from target documents with respect to their corresponding source documents and develop a classification system. Our results are promising and can serve as robust baseline to further research in this topic.

[1]http://www.iitp.ac.in/∼ai-nlp-ml/resources.html

2. Related Works

Although sentence level novelty detection is a well studied problem in information retrieval literature, very little has been done to address the problem at the document level. To begin with (Li and Croft, 2005) rightly pointed out that, research in novelty detection from texts has been carried out at three levels : event level, sentence level and document level. Research in novelty mining could be traced back to the Topic Detection and Tracking (TDT) (Allan, 2002) evaluation campaigns where the concern was to detect new event from online news streams. Although the intention was to detect the *first story* or reporting of a new event from a series of news stories, the notion of *novelty detection* from texts came into light for the research community. Some notable approaches for New Event Detection with the TDT corpus are by (Allan et al., 1998; Yang et al., 2002; Stokes and Carthy, 2001; Franz et al., 2001; Yang et al., 1998; Allan et al., 2000; Brants et al., 2003). However, the Novelty track in Text Retrieval and Evaluation Conferences (TREC) (Soboroff and Harman, 2005) were the first to explicitly explore the concept of Novelty Detection from texts. Under the paradigm of information retrieval, given a query, the TREC experiments were designed to retrieve relevant and novel sentences from a given collection. Some notable approaches for sentence level novelty detection from the TREC exercises are by (Allan et al., 2003; Kwee et al., 2009; Li and Croft, 2005; Zhang et al., 2003; Collins-Thompson et al., 2002; Gabrilovich et al., 2004; Ru et al., 2004). Textual Entailment based sentence level novelty mining was explored in the novelty subtask of Recognizing Textual Entailment-Text Analytics Conference (RTE-TAC) 6 and 7 (Bentivogli et al., 2011). The datasets made available from these tasks are for sentence level novelty mining and were created from an information retrieval perspective. At the document level the problem is attempted by a few like (Zhang et al., 2002; Tsai and Zhang, 2011; Karkali et al., 2013; Dasgupta and Dey, 2016). However none of the datasets used in these works are publicly available. Hence we find that there is a dearth of a proper evaluation setup (e.g. corpus, baseline and evaluation methods) for document level novelty detection. This inspired us to create one and establish a benchmark for the same.

3. Motivation and Contribution

Our understanding and survey revealed that in spite of having several applications in NLP tasks, novelty detection at the document level has not attracted the coveted attention. Hence, we deem that novelty at the document level needs to be understood first, investigated in-depth, and benchmark setup (gold standard resources) be created to validate the investigations. We hope that the knowledge gained from this dataset and experiments would be a step towards our more ambitious vision of semantic level plagiarism detection in scholarly articles. We briefly outline the contributions of this work :

- Proposing a benchmark dataset for document level novelty detection. We are unaware of the availability of any such corpus; and
- Presenting a supervised machine learning model for document level novelty detection. This can be treated as a baseline model for further research along this line.

4. Document Level Novelty

Novelty detection from texts implies search for new information with respect to whatever is already known or seen. Hence, the problem of novelty detection from texts is very subjective and depends upon the view of the intended reader. The knowledge of the reader regarding a particular event serves as the reference against which s/he decides the novelty of an incoming information. Careful observation of data characteristics led us to believe that there exists at least four properties that characterizes novelty detection from texts :

- *Relevance*
- *Diversity*
- *Relativity* and
- *Temporality*

4.1. Relevance

The target document should be relevant to prior knowledge. For example, seeking novelty between two documents, one talking about *jaguar*, the animal and the other about *jaguar*, the car is futile as one is not relevant to the other. Quite obvious that each one would contain different information than the other. So *Relevance* should hold.

4.2. Diversity

Diversity correlates with the new information content. More the new information in a document, diverse would be the content. Hence novel information should be relevant yet *diverse* from existing information.
For example, let us consider, on a given date a certain newswire document X reports about an accident at a certain place. On the subsequent date another reporting X' surfaces which details about the investigation being carried out by police. Now X' will contain new information with respect to X. That is to say, given a reader has already read about the first reporting (facts of the accident) X, the second reporting X' having significant different content as well as different direction of reporting (or intent) would appear novel to the reader. So X' is relevant to X yet divergent i.e. containing new information.

4.3. Relativity

The amount of new information content is important while deciding the novelty of an entire document. When we talk about a document being novel it is always with respect to a reference set of documents already seen (information already gained from those seen documents) or what we say as the knowledge base of the reader. So the quantity of *relative* new information plays a role for deciding document novelty.

4.4. Temporality

Finally novel information is usually a temporal update over existing knowledge. The previous example justifies the view.

With these notion of novelty we went on to create a resource that effectively taps these properties, *viz.,* **Relevance**, **Diversity**, **Relativity** and **Temporality**. Our resource not only encompasses the lexical form of redundancy (a straight forward form of *non-novelty*) but also delves deep into semantic textual redundancy (a more complex form of *non-novelty*). To understand it better let us consider the following example :

- *d1 : Singapore is an island city-state located at the southern tip of the Malay Peninsula. It lies 137 kilometers north of the equator.*
- *d2 : Singapore's territory consists of one main island along with 62 other islets. The population in Singapore is approximately 5.6 million.*
- *d3 : Singapore is a global commerce, finance and transport hub. Singapore has a tropical rain forest climate with no distinctive seasons, uniform temperature and pressure, high humidity, and abundant rainfall.*
- *d4 : Singapore, an island city-state off southern Malaysia, lies one degree north of the equator. As of June 2017, the island's population stood at 5.61 million.*

It is fairly easy to conclude that *d4* follows from *d1* and *d2*. However, considering only *d3* as the source, although related but *d4* has entirely diverse information. Thus, *d4* would be non-novel with respect to *d1* and *d2* but would appear novel with respect to *d3* only. We build our corpus along this line and the annotations too followed this kind of judgments.

5. Benchmark Setup : TAP-DLND 1.0

To address the issues pointed out in the previous section we develop a benchmark setup as discussed below.

5.1. Data Collection

We design a web crawler[2] to perform systematic, unbiased, *event-specific* crawling of news articles, mostly from the online versions of Indian English newspapers. The news domains we looked into are : *Accident (ACC), Politics (PLT), Business (BUS), Arts and Entertainment (ART), Crime (CRM), Nature (NAT), Terrorism (TER), Government (GOV), Sports (SPT), and Society (SOC)*. To ensure that *Temporality* criteria is preserved, our web crawler is designed to fetch web documents for a certain event in a timely manner i.e. the crawled documents are grouped as per their dates of publications in different forums (See Figure 1). Event wise statistics of the corpus are in Table 2.

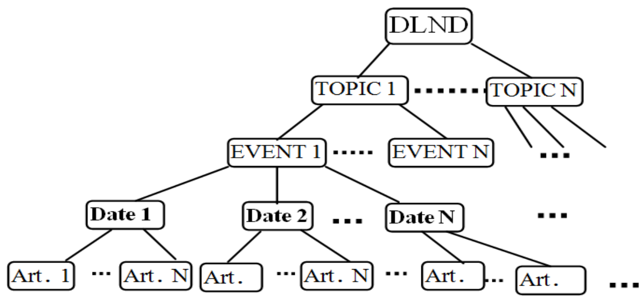

Figure 1: Temporal Crawling

Features	Statistics
Crawling period	Nov'16 - Nov'17
Number of events	223
Number of sources per event	3
Total novel documents	2736
Total non-novel documents	2704
Total documents in TAP-DLND 1.0	6109
Average number of sentences	15
Average number of words	353

Table 1: Statistics of TAP-DLND 1.0 corpus. Here, average number of sentences and words is per document.

5.2. Preprocessing

As the data were crawled from various web sources we perform some manual preprocessing works like removal of headlines, news source, date, time, noises (advertisements, images, hyperlinks) and convert the data into desired shape.

[2]using the www.webhose.io API

Category	# Events	# N	# NN
ACC	10	231	272
PLT	97	669	685
BUS	35	202	264
ART	21	397	258
CRM	10	237	174
NAT	10	87	250
TER	18	255	468
GOV	15	405	219
SPT	2	39	51
SOC	5	214	63

Table 2: Event wise statistics of TAP-DLND 1.0, $\#N \rightarrow$ Number of Novel documents, $\#NN \rightarrow$ Number of Non-Novel documents

5.3. Source Document Selection

To mandate the *Relevance* and *Relativity* criteria, we select three documents for each event as the seed source documents. They are usually selected from the initial dates of reporting of a particular event. Also so chosen that they represent different facets of information regarding that particular event (*information coverage*). These source documents serve as the reference against which we asked the annotators to tag a target document (chosen from the remaining crawled documents for that event) as *novel* or *non-novel*. The source documents could be perceived as the memory of the reader or information already known against which it is to be determined with reasonable level of certainty that whether a target document contains sufficient new information to be labeled as *novel*.

5.4. Renaming files

For ease of information retrieval we rename each document in the corpus. A certain document bearing **ACCE005SRC003.txt** as file name indicates that it is the 3rd source document of the 5th event in the *accident category*. For target documents SRC is replaced by TGT.

5.5. Meta files

We generate meta files (.xml) for each document in the corpus. These meta files contain background information regarding a source/target document within structured XML tags and have the same file name as that of the corresponding document. The information content of the meta files are : *date of publishing, publisher, title of reporting, source id, event id, event name, category, Document Level Annotation (DLA), number of words and sentences*. We develop a semi-automatic *meta file generator interface* where attribute values are automatically captured from the hierarchically organized data (See Figure 2). Stanford CoreNLP (Manning et al., 2014) integrated with our interface gave us the field values for *sentence* and *word* count. We asked our annotators to provide their judgments for the *DLA* attribute based on the guidelines specified in the next section.

5.6. Annotation

Three annotators with post-graduate level knowledge in English were involved in labeling the TAP-DLND 1.0 target documents. Having read the source document(s) we asked the annotators to annotate an incoming on-event document as *non-novel* or *novel* solely based on the information coverage in the source documents. The **annotation guidelines** were simple:

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. To leave out the ambiguous cases (for which the human annotators were not sure about the label).

Two annotators independently labeled the target documents. The third annotator resolved the differences via majority voting. We found that novel items with respect to the source documents were mostly found in the reporting published in subsequent dates. Whereas non-novel items we found in the reporting published by different agencies in the same date as that of the source documents. This is in line with the *Temporality* criteria we discussed earlier. The inter-annotator agreement ratio was found to be **0.82** in terms of **Kappa coefficient** (Fleiss, 1971) which is assumed to be good as per (Landis and Koch, 1977). The final structure of DLND is in Figure 2.

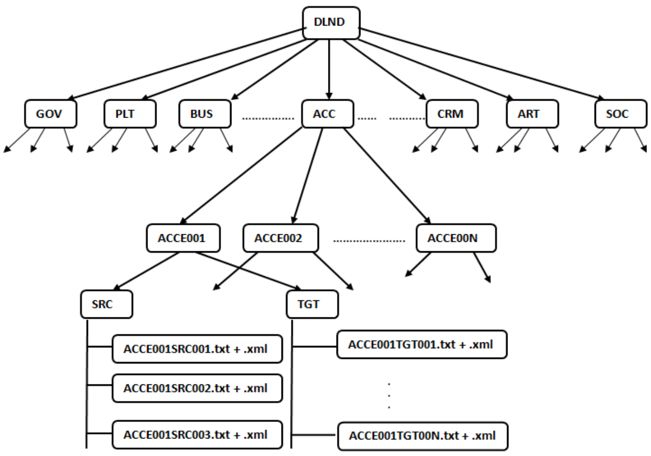

Figure 2: The DLND corpus structure

6. Evaluation

We frame document level novelty detection as a binary classification problem and choose the features in parlance with the objective nature of texts that we consider for our experiments. We develop a binary classifier based on Random Forest[8] (RF) (Breiman, 2001) algorithm that classifies a document into either *novel* or *non-novel*. Our key focus is on extracting features that contribute to the semantics of a document. The set of features that we use for training and/or testing RF is listed in Table 3. As is evident from the discussion in Section 3, TAP-DLND 1.0 consists a fair share of different levels (lexical as well as semantic) of text representations.

6.1. Results and Discussions

We first take a simple yet popular lexical baseline: Jaccard similarity with unigrams between the source document and the target (Zhang et al., 2003). We train a Logistic Regression (LR) classifier with the Jaccard score to classify a document based on its overlap with the source document. Table 4 clearly indicates that the lexical baseline fails miserably in identifying *non-novel* documents. Next we went ahead with three approaches by (Zhang et al., 2002) for novelty detection at the document level. The first one i.e. the Set Difference is essentially the count of new words in the target document with respect to the set of source document(s). For this we concatenate the source document(s) of each event to form one source against each target. The Geometric Distance measures the cosine similarity between two document vectors represented as *tf-idf* vectors. For three source documents against one target document in TAP-DLND 1.0, we take the maximum of the cosine similarity score. The third approach measures the Kullback-Leibler divergence between the concatenated source document(s) and the prospective target document where a document d is represented as a probabilistic unigram word distribution (language model θ_d). Instead of setting a fixed threshold as (Zhang et al., 2003), we train a Logistic Regression classifier based on those measures to automatically determine the decision boundary. Another approach by (Karkali et al., 2013) based on Novelty Scoring via Inverse Document Frequency (IDF) performed poorly in recognizing novel/non-novel documents in TAP-DLND 1.0. We also compare our method with a more recent approach of (Dasgupta and Dey, 2016) on our data. This particular entropy-based approach produces novelty score (NS) of a document d with respect to a collection C. We adapt their respective threshold criteria and infer that documents with novelty score above (*average+standard deviation*) are *novel* and that with novelty score below (*average-standard deviation*) are *non-novel*. We left out the remaining (average novelty class) cases for our experiments. Table 4 numbers clearly show that our method superseded the baselines and purported *state-of-the-art* by a substantial margin.

6.2. Feature Significance

We investigated the significance of each feature by measuring the Information Gain (See Figure 3). The information gain for a feature x_k is the expected reduction in entropy-i.e., uncertainty achieved by learning the state of that feature. We attribute the better performance of our approach

[3]Distributed Bag-Of-Words (DBOW) paragraph vector model trained on Wikipedia articles.

[4]Trained on Google News Corpus of 100 billion words. 300 dimension vectors using CBOW model

[5]Entities were extracted using the Stanford Tagger.

[6]Using the Rapid Automatic Keyword Extraction (RAKE) algorithm.

[7]Obtained from English WordNet (Fellbaum, 1998)

[8]RF of 100 trees with minimum number of instances per leaf set to 1 implemented in WEKA machine learning toolkit

Type	Features	Description
Semantic	Paragraph Vector (pv) + Cosine	We represent the source and target documents in terms of *paragraph vectors*[3](Le and Mikolov, 2014). Then we take the maximum of the cosine similarity between the source-target pairs.
Semantic	Concept Centrality	To identify the central theme of a document we use the *TextRank* summarization algorithm by (Mihalcea and Tarau, 2004). Thereafter we vectorize the ranked summary for each source and target document by simple *word2vec*[4](Mikolov et al., 2013) concatenation. Finally we take the maximum of the cosine similarity between the source and target vectors.
Lexical	n-gram similarity	We compute lexical overlap of target *n-gram*'s with respect to source documents for n = 2,3 and 8. Octagrams we use to put emphasis on phrase overlap.
Lexical	Named Entities and Keywords match (kw-ner)	As Named Entities[5] and Keywords[6] play a significant role in determining *relevance*, we put additional weightage to them by considering their match (target w.r.t. sources) as a separate feature.
Lexico-Semantic	New Word Count (nwc)	The number of new words could be an effective indicator of the amount of novel information content in the target document w.r.t. the source(s) given. Here, for calculating new words, along with the surface forms, we consider their synonyms[7] as well to establish semantic relatedness.
Language Model	Divergence (kld)	We use this feature to measure the dissimilarity between two documents represented as language models. We concatenate all the source documents into one and then measure the Kullback-Leibler Divergence with the target.

Table 3: Feature Set

Systems	P(N)	R(N)	F_1(N)	P(NN)	R(NN)	F_1(NN)	Accuracy
Jaccard+LR (Baseline)	52.2	96.1	67.6	74.0	10.9	19.0	53.8
Set Difference+LR (Zhang et al., 2002)	74.3	71.5	72.8	72.2	74.9	73.5	73.2
Geometric Distance+LR (Zhang et al., 2002)	65.6	84.3	73.7	84.2	55.3	66.7	69.8
Language Model (KLD)+LR (Zhang et al., 2002)	73.2	74.9	74.1	74.0	72.3	73.1	73.6
Novelty (IDF)+LR (Karkali et al., 2013)	52.5	92.1	66.9	66.5	15.9	25.6	54.2
(Dasgupta and Dey, 2016)	65.1	63.8	64.4	64.1	65.3	64.6	64.5
Proposed Approach (RF)	**77.6**	82.3	**79.8**	80.9	**76.1**	**78.4**	79.2

Table 4: *10-fold* cross-validation results on TAP-DLND 1.0 (in %), $P \rightarrow Precision$, $R \rightarrow Recall$, $N \rightarrow Novel$, $NN \rightarrow Non-Novel$, $LR \rightarrow$ Logistic Regression, $IDF \rightarrow$ Inverse Document Frequency, $KLD \rightarrow$ Kullback- Leibler Divergence

to the choice of semantic features for our experiments (see Figure 3). Lexico-Semantic feature new word count has the maximum contribution, for which we argue that novel events in context to newspaper articles would contain new entities, concepts, numbers whereas non-novel documents would consist identical or synonymous entities. Semantic features play a vital role which indicates that detection of novelty extends beyond lexical characteristics of text.

7. Conclusion

In this work we put forward a benchmark resource for *document level novelty detection* and an evaluation scheme for the same. Our resource has an extensive coverage of ten different news categories and also includes the *relevance, diversity, relativity,* and *temporality* criteria inherently within its schema. Along with straightforward lexical characteristics it also manifests the high level semantic understanding of human annotators in its gold labels which is very essential for detecting semantic level redundancy. We hope that TAP-DLND 1.0 would evolve as a benchmark resource for experiments on document level novelty detection and provide valuable insights into the problem. In future we plan to annotate the TAP-DLND 1.0 corpus at the sentence level to have more fine perception regarding the amount of new information required to deem a document as *novel*. Also we intend to include more target documents in data scarce categories.

8. Availability

The dataset is submitted to European Language Resource Association (ELRA) for hosting. Also would be available at http://www.iitp.ac.in/∼ai-nlp-ml/resources.html. Researchers can also contact the first author for the resource.

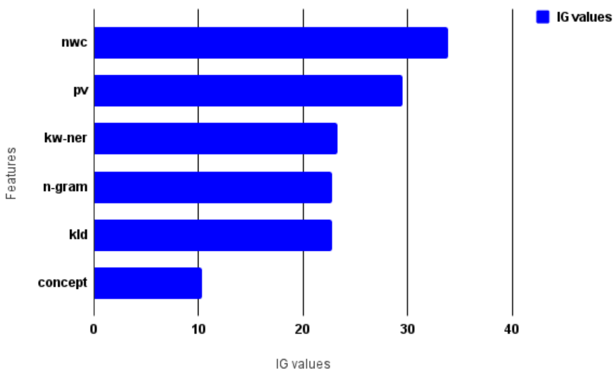

Figure 3: Significance of features *based on Information Gain (IG)*. The length of the bar corresponds to the average merit (X : IG) of the feature (: Y).

9. Acknowledgements

Tirthankar Ghosal and Asif Ekbal gratefully acknowledge Visvesvaraya PhD scheme for Electronics and Information Technology, an initiative of Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia). Tirthankar is a PhD scholar under Visvesvaraya PhD scheme and Asif is the recipient of Sir Visvesvaraya Young Faculty Research Fellow Award. The work is also generously supported by Elsevier Centre of Excellence for Natural Language Processing, Department of Computer Science and Engineering, Indian Institute of Technology Patna. We would also like to thank the anonymous reviewers for their valuable suggestions in improving this work.

10. Bibliographical References

Allan, J., Papka, R., and Lavrenko, V. (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. ACM.

Allan, J., Lavrenko, V., and Jin, H. (2000). First story detection in tdt is hard. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 374–381. ACM.

Allan, J., Wade, C., and Bolivar, A. (2003). 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.

Allan, J. (2002). Introduction to topic detection and tracking. *Topic detection and tracking*, pages 1–16.

Bentivogli, L., Clark, P., Dagan, I., and Giampiccolo, D. (2011). The seventh pascal recognizing textual entailment challenge. In *TAC*.

Brants, T., Chen, F., and Farahat, A. (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. ACM.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Collins-Thompson, K., Ogilvie, P., Zhang, Y., and Callan, J. (2002). Information filtering, novelty detection, and named-page finding. In *TREC*.

Dasgupta, T. and Dey, L. (2016). Automatic scoring for innovativeness of textual ideas. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

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

Franz, M., Ittycheriah, A., McCarley, J. S., and Ward, T. (2001). First story detection: Combining similarity and novelty based approaches. In *Topic Detection and Tracking Workshop Report*, pages 193–206.

Gabrilovich, E., Dumais, S., and Horvitz, E. (2004). Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th international conference on World Wide Web*, pages 482–490. ACM.

Karkali, M., Rousseau, F., Ntoulas, A., and Vazirgiannis, M. (2013). Efficient online novelty detection in news streams. In *WISE (1)*, pages 57–71.

Kwee, A. T., Tsai, F. S., and Tang, W. (2009). Sentence-level novelty detection in english and malay. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 40–51. Springer.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196.

Li, X. and Croft, W. B. (2005). Novelty detection based on sentence level patterns. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 744–751. ACM.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into texts. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Ru, L., Zhao, L., Zhang, M., and Ma, S. (2004). Improved feature selection and redundance computing-thuir at trec 2004 novelty track. In *TREC*.

Soboroff, I. and Harman, D. (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. Association for Computational Linguistics.

Stokes, N. and Carthy, J. (2001). First story detection using a composite document representation. In *Proceed-*

ings of the first international conference on Human language technology research, pages 1–8. Association for Computational Linguistics.

Tsai, F. S. and Zhang, Y. (2011). D2s: Document-to-sentence framework for novelty detection. *Knowledge and information systems*, 29(2):419–433.

Yang, Y., Pierce, T., and Carbonell, J. (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. ACM.

Yang, Y., Zhang, J., Carbonell, J., and Jin, C. (2002). Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 688–693. ACM.

Zhang, Y., Callan, J., and Minka, T. (2002). 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. ACM.

Zhang, M., Song, R., Lin, C., Ma, S., Jiang, Z., Jin, Y., Liu, Y., Zhao, L., and Ma, S. (2003). Expansion-based technologies in finding relevant and new information: Thu trec 2002: Novelty track experiments. *NIST SPECIAL PUBLICATION SP*, (251):586–590.