applied sciences

MDPI

Article

Linked Data Triples Enhance Document Relevance Classification

Dinesh Nagumothu *, Peter W. Eklund, Bahadorreza Ofoghi, Mohamed Reda Bouadjenek

School of Information Technology, Deakin University, Geelong, Victoria 3220, Australia; peter.eklund@deakin.edu.au (P.W.E.); b.ofoghi@deakin.edu.au (B.O.); reda.bouadjenek@deakin.edu.au (M.R.B.)
* Correspondence: dnagumot@deakin.edu.au

Abstract: Standardized approaches to relevance classification in information retrieval use generative statistical models to identify the presence or absence of certain topics that might make a document relevant to the searcher. These approaches have been used to better predict relevance on the basis of what the document is "about", rather than a simple-minded analysis of the bag of words contained within the document. In more recent times, this idea has been extended by using pre-trained deep learning models and text representations, such as GloVe or BERT. These use an external corpus as a knowledge-base that conditions the model to help predict what a document is about. This paper adopts a hybrid approach that leverages the structure of knowledge embedded in a corpus. In particular, the paper reports on experiments where linked data triples (subject-predicate-object), constructed from natural language elements are derived from deep learning. These are evaluated as additional latent semantic features for a relevant document classifier in a customized news-feed website. The research is a synthesis of current thinking in deep learning models in NLP and information retrieval and the predicate structure used in semantic web research. Our experiments indicate that linked data triples increased the F-score of the baseline GloVe representations by 6% and show significant improvement over state-of-the art models, like BERT. The findings are tested and empirically validated on an experimental dataset and on two standardized pre-classified news sources, namely the REUTERS and 20 NEWSGROUPS datasets.

Keywords: linked data triples; named entities; topic modeling; deep learning; relevance classification

Citation: Nagumothu, D.; Eklund, P.W.; Ofoghi, B.; Bouadjenek, M.R. Linked Data Triples Enhance Document Relevance Classification. *Appl. Sci.* **2021**, *1*, 0. https://doi.org/

Academic Editor: Sławomir Nowaczyk, Mohamed-Rafik Bouguelia, Hadi Fanaee

Received: 16 June 2021
Accepted: 16 July 2021
Published:

1. Introduction

Pre-classification of news-feed articles based on user feedback can play an important role in preventing users being overwhelmed with information. In the context of social media, friend importance and textual features are utilized [1] to classify news-feed items on Facebook achieving 64% accuracy using a Support Vector Machine (SVM) classifer. In addition, on Facebook, work has been conducted on classification of feeds on user walls to distinguish between friend posts and liked pages [2], with further sentiment analysis on life-event-related friend posts. A similar document classification and sentiment analysis of news-feeds is conducted in Reference [3] using Bayesian and SVM classifiers. More recently, classification of news articles has been approached using both dictionary-based and machine learning-based techniques on features extracted from sentences represented as text segments (rather than full text) of news articles [4]. Robuzz [5], a New York-based start-up specializing in smart news-feeds, makes use of deep learning techniques and textual features of news content to provide users with a relevant news-feed.

While previous work on classification of news-feeds and articles has been mostly around the utilization of bag-of-words and other flat textual features, in this work, we focus on the adaptive analysis of news-feed items from a semantic web perspective, namely using structural elements developed for the semantic web as additional features. Our techniques facilitate the development of a focused crawler that predicts the relevance of documents arriving to a news-feed using several lexical semantic features extracted from the content

of documents already accumulated. For this purpose, we extract and utilize the major Named Entities (NEs) in the news articles—persons, locations, organizations, products, etc.—model and find abstract "topics" within each article, and also represent and encode three-way lexical relationships in chunks of texts in the form of subject-predicate-object, referred to as triples or linked data (https://en.wikipedia.org/wiki/Linked_data (accessed on 21 January 2021)).

There is evidence in the prior literature that topics, i.e., extracted using the Latent Dirichlet Allocation (LDA) technique [6]—a type of statistical modeling for discovering the abstract topics—, in combination with Named Entities (NEs) are features that can semantically enrich documents [7]. Thus, adding LDA topics and NEs is hypothesized to have a positive effect on the performance of a news-feed relevance classifier. On the other hand, the structural features of text, namely the triples, constructed based on the linked data model developed for the semantic web, can capture and encapsulate semantically related chunks of text within each news article.

We hypothesize that the utilization of linked data triples that contain LDA topics and NEs and will show that these features provide additional latent semantic information to the task of textual relevance classification in news-feeds. Thus, the contributions of this research are as follows: (i) we confirm the results obtained by Kim et al. [7], namely that LDA topics and NEs, when added as additional features to a corpus, semantically enrich the corpus. This affirms what we refer to as the *topic-entity model*, a model in which LDA topics and named entities complement the original document corpus; (ii) we further show that additional structural features, linked data triples formed by LDA topics and NEs at the sentence level, have an even greater positive latent semantic effect on news-feed classification, we call this the *topic-entity-triple* model.

The contributions of this work are as follows:

A binary relevance classification dataset: named ENERGY-HUB, containing news articles from various publishers collected through Google NEWS API.
Topic-Entity Triples: a novel way to generate triples from raw text with topics and named entities is proposed.
Semantic enrichment to enhance document representations using linked-data triples: we conduct several experiments on a dataset of our domain focus, the ENERGY-HUB news articles, and then validate our findings on two separate standard datasets, namely the REUTERS (http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html, accessed on 5 Feb 2021) and 20 NEWSGROUPS (https://www.kaggle.com/crawford/20-newsgroups, accessed on 5 Feb 2021) datasets. Results indicate that documents enriched with linked-data triples improve classification performance over and above those without enrichment.

The paper is organized as follows: Section 2 introduces existing literature on document enrichment with semantic features, and Section 3 provides a discussion on natural language (NL) features extracted for enrichment; the ENERGY-HUB section (Section 4) describes an experimental framework with data source, as well as its corpus; Learning Models (Section 5) discusses the proposed architecture; Experiments (Section 6) presents the experimental results against in-house ENERGY-HUB dataset and standard REUTER'S 21,578 and 20 NEWS GROUPS datasets. Discussion (Section 7) provides an overall discussion of the results, including future work. Concluding Remarks (Section 8) provides conclusions and final comments.

2. Related Work

There is a substantial research related to enhancing document representation using external source of data. In Figure 1, we summarize the main techniques used for that goal, namely word embeddings and ontology- or knowledge-based approaches.

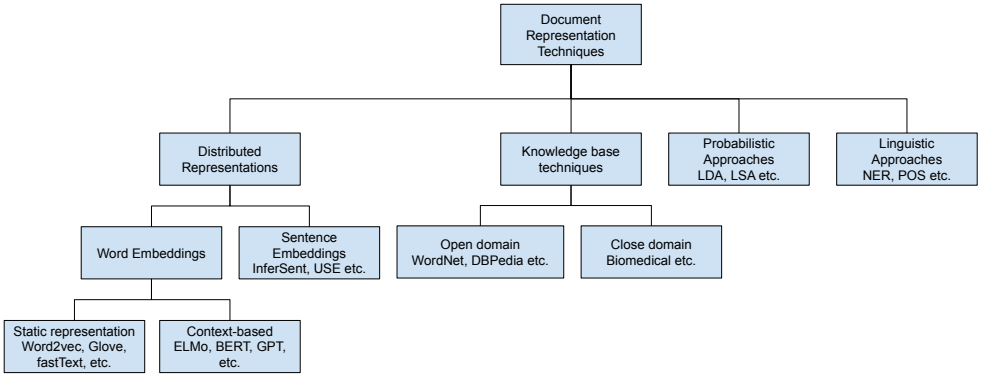

Figure 1. A taxonomy of document representation enhancement techniques.

Feature extraction plays an important role in representing documents as vectors in document classification. Initially, approaches used a bag-of-words model with term frequencies and inverse document frequencies as features. There have been numerous advances in recent years achieving more effective representations of textual documents. Word embeddings, such as Word2vec [8], GloVe [9], and fastText [10], provide a fixed low-dimensional vector representation for every representative term within a document. Sentence embeddings, like InferSent [11] and Universal Sentence Encoder [12] (USE), provide fixed vector representations for every sentence in the document. The introduction of language-models, such as ELMo (Embeddings from Language Models) [13] and Bi-directional Encoder Representation of Transformers [14], has resulted in further advances, by enhancing the contextualized distributed of the representations of terms.

Enriching document representations using semantic features, as in this work, is aimed at improving relevance classification performance. For instance, in Reference [15], enhanced document representations are formed by embedding background knowledge from a relevant domain ontology. The terms in documents are matched against the concepts in the ontology. If an exact/partial match is found, the terms are replaced by the associated concepts forming a richer semantic representation of the document. In Reference [16], document related concepts are fetched from the domain ontology forming a CF-IDF (Concept Frequency-Inverse Document Frequency) to classify news-feeds. In Reference [17], news documents are clustered by measuring similarity between named entities from a document and related matched information from DBPedia (https://www.dbpedia.org/, accessed on 10 May 2021.) A Semantic Enriched Deep Learning Architecture (SE-DLA) is proposed in Reference [18], this approach uses semantic vectors from WordNet to represent documents. WordNet [19] is a widely used resource containing taxonomy of words with their semantic relations. These semantic vectors are fused with general word-embeddings to form a vector space representation which is then fed into a neural network to obtain output relevance classes. In Reference [20], semantic enrichment is performed with external knowledge bases (KBs), for short text classification. Relevant concepts are retrieved from knowledge-bases, such as YAGO [21] and Freebase [22], and embedded into the model via concatenation. Documents belonging to a particular class are often linked together, hence providing better representation. Considering this assumption, a new document embedding method is proposed in Reference [23] using the hyperlinks and relations between documents for classification. Open domain knowledge bases are less useful in complex domains, like medical document classification [24]. Incorporating domain knowledge with relevant closed-domain ontologies are used in these situations to enrich document representations. However, the main weakness of these approaches is the reliance on external ontologies to improve the semantic richness of the document.

The selection of ontology is domain-specific and needs to be relevant to the corpus. Open domain knowledge-bases may not be an ideal solution as they suffer incompleteness [20]. In contrast, our work utilizes linked-data triples (formed using natural language

features—abstract topics and named entities) generated entirely from within the corpus itself, without recourse to external ontologies or knowledge sources.

Probabilistic approaches, like Topic Models, use corpus-word distributions to categorize documents into *n* different topics. Each document is represented as a probability distribution of topics, and each topic is represented as probability distribution of words in the corpus. Topic models can assist in semi-supervised document classification (with minimal labeled data) [25]. Initially, labeled and unlabeled sets are combined to represent documents with topics. Self-learning is performed by computing similarities between documents from labeled and unlabeled sets to find the best suitable label for each document. Finally, this combined labeled set is used to train a supervised classifier to make predictions on unseen data.

Multi-modal architectures help to fuse different features into a single model. In Reference [26], images and their textual descriptions are combined to perform sentiment analysis. Image and text-based models are created separately and fused to perform multi-label sentiment classification. In our work, multiple features, like text, topic distributions, named entities, and triples, are fused using layer concatenation for classifying document relevance.

Regarding the taxonomy of document representational approaches in Figure 1, our approach is hybrid. *Named-entity extraction* is an example of *Linguistic enrichment (Language dependent)*, so it is an *open domain* ontology technique in Figure 1. *Latent Dirichlet Allocation (LDA)* topic modeling is an example of a *Probabilistic approach* in Figure 1. *Linguistic enrichment* is also at play with the formation of Linked data triples by analyzing the syntactic dependencies of the document sentences. On the other hand, we also deploy *Word embeddings*, considering the linked data triples as primitive sentences, and we are able to leverage *context-based word embeddings*, comparing the performance of different methods from that domain, namely GloVe, Infersent, and BERT, to baseline document relevance classification.

3. Natural Language Features

3.1. Topic Models

Topic modeling is a natural language processing (NLP)-based text mining technique used to find hidden semantic structure in the form of abstract topics from a corpus of documents by analyzing word distributions. It is predominantly used for document retrieval [27], text classification [28], text summarization [29], and relevance mapping between documents [30]. Latent Dirichlet Allocation (LDA) [6] is the most popular topic modeling technique to discover abstract topics present within a document collection. In LDA, each document is represented as a probabilistic distribution over latent topics and each topic as a distribution of words. For each document, topic distributions and keywords are learned from LDA modeler. Topic distributions hold the probabilities of topics that the document is "about". The top-n terms of the topics with highest probabilities are considered as topic-keywords.

3.2. Named Entities

Named Entity Recognition (NER) of Named Entities (NEs) is an information extraction technique that identifies named and numeric entities from text—persons, locations, organizations, products, etc. NER has been previously achieved using lexicons, rule-based techniques, and ontologies [31], as well as with neural networks and machine learning techniques [32].

3.3. Linked Data (triples)

Triples (or linked data) are generally of the form *subject-predicate-object* triples and multiple triples can be extracted from every complete sentence in a document. As a document contains a set of sentences, each document generates a set of triples. Triples from a sentence [33] can be extracted by syntactic parsers, such as OpenNLP (https://

opennlp.apache.org/, accessed on 8 March 2021), Stanford parser [34], Link Parser (https://www.link.cs.cmu.edu/link/, accessed on 8 March 2021), and Minipar [35]. All of these systems generate similar outputs; however, their extraction algorithms may differ. In the Stanford Parser, for instance, each sentence is given as input, and the parser produces a syntactical Treebank-style [34] output. A Treebank defines the syntactical structure of a sentence. A sentence *S* occupies the top of the tree as its root node. A sentence is split into three children—Noun Phrase (NP), Verb-Phrase (VP), and full stop. The last child, full-stop, does not contain any useful information but serves simply to syntactically separate sentences. The other two phrases have part-of-speech tags on each word.

Triple extraction proceeds in three phases: finding the subject, the predicate, and the object of the sentences from the NPs and VPs. To find the subject, a breadth-first search algorithm is implemented over the NPs to identify the first noun in the phrase. Nouns are usually tagged as NN (Common noun), NNP (Singular proper noun), NNPS (Plural, proper noun), and NNS (Plural common noun). To determine the predicate, a search function is conducted on the VPs, and the deepest verb is extracted and considered. Verbs are tagged as VB (base verb), VBD (past tense verb), VBG (present participle verb), VBN (past participle verb), VBP (present tense first person singular), and VBZ (third person singular). Lastly, objects can also be fetched from VPs but should be searched in three sub-phrases—NPs, Adjective-Phrases (ADJPs), and Prepositional-Phrases (PPs). Nouns are searched in NPs and PPs for object extraction, while adjectives in ADJPs are considered as attributes of the object. JJs (ordinal adjectives) and JJRs (comparative adjectives) are tags that belong to ADJPs.

For instance, consider a simple sentence in a document that reads *"U.S. scientists call for fight against climate change"*. *"U.S. scientists"* is the noun phrase, hence being considered the subject. In VP, *"call"* being the only verb is tagged as the predicate. This sentence has two noun phrases, hence forming two objects *"fight"* and *"climate change"*. The words *"for"* and *"against"* in the sentence are adjective phrases (ADJP), considered as attributes to the object. The two extracted linked data triples corresponding to the utterance *"U.S. scientists call for fight against climate change"* would be <U.S. scientists, call, for fight> and <U.S. scientists, call, for climate change>.

4. ENERGY-HUB

The news-feed we have developed for our research client concerns energy policy, hence the name ENERGY-HUB; however, the same methods can be applied in other domains. The underlying architecture of the system consists of an app engine, a topic modeler, a text classifier, a Named Entity recognition system, a NoSQL database, and a corpus of crawled documents. The system initializes itself with the reader making search requests from a home page and the presentation of the site, its affordances, and navigational elements are intended to adapt to the person using the site over time. The reader can search for other articles, as well, with the default search affordances, e.g., a user might indicate an interest in articles in business newspapers on renewable energy, particularly hydrogen production, from Europe and North America.

Once a sufficient corpus had been accumulated in the ENERGY-HUB against the user's search, the articles in the corpus are tagged by domain experts as relevant/irrelevant. This dataset forms the basis for our training set. The relevant set of documents in the corpus representing positive training instances, the irrelevant representing negative training instances. The very same corpus can now be used as a training set for a classifier to predict the relevance/irrelevance of new arrivals in the news-feed. In this work, about 2751 documents represent a critical mass of articles in the corpus used to train a neural network-based classifier; this is further discussed in Learning Models section (Section 5).

4.1. Data Source

The news API (https://newsapi.org/s/google-news-api, accessed on 21 January 2021) from Google acts as the principal news-feed input. For every search request, the API

responds with a list of articles. The result-set holds metadata about each article, including "title", "URL", "source", "author name", and "article date". We use "energy + [country name]" as the initial search terms, where [country name] = {"Australia" or "United States of America" or "India" or "Canada" or "Germany" or "United Kingdom" or "New Zealand"}, and the country list can be altered as required. These search terms initialize the news-feed and provide a stream of articles to the site arriving at regular intervals. Considering reader preferences, articles from business newspapers (https://en.wikipedia.org/wiki/List_of_business_newspapers, accessed on 21 January 2021) only are selected from the North American and European sources and the additional keywords *"sustainable energy"* and *"hydrogen production"* provide additional filters on the feed.

After gathering article meta-data through the Google news API, each article is fetched using "urllib3" (a Python-based HTTP client). The textual data is scraped by identifying all tags using Beautiful Soup, a Python package deployed on top of a HTML parser. The articles are then pipe-lined to be stored in a NoSQL database, MongoDB. This process is run at regular intervals to update the ENERGY-HUB content frequently. This approach can result in the retrieval of duplicate articles; this is avoided by indexing the URL of the collection and tagging it as unique.

4.2. Corpus

The purpose of the ENERGY-HUB news-feed is to build a corpus of articles about policy-making, production, transmission, and consumption of energy across the globe with minimal editorial or moderator support. Since we use a generic search API, news items irrelevant to energy policy can inadvertently match the search criteria and deteriorate the performance of the adaptive user interface through their inclusion. For example, an irrelevant article about the "Women's cricket world cup" (https://www.abc.net.au/news/2020-02-29/jess-jonassen-on-the-t20-world-cup-and-australias-campaign/12010560, accessed on 22 January 2021), found in the news API results, is mistakenly considered energy-related by virtue of the sentence*"Molly Strano has come in for Tayla and she has added a bit of energy and positivity to the squad too."*

The corpus currently has 2751 articles, of which 1078 items are identified as irrelevant. Each document has been judged as relevant/irrelevant by domain experts. The dataset is available at: https://github.com/dineshnagumothu/energyhubnews (accessed on 06 July 2021).

The operation of the site, as described, results in a tagged corpus—a collection of relevant and irrelevant documents—which accumulated over several months. The corpus contains 1673 relevant and 1078 irrelevant articles. This gives rise to the idea of using this collection to train a neural network-based classifier to automatically filter irrelevant articles from the ENERGY-HUB news-feed. We call the interaction between the news-feed and the neural network classifier "a focused crawler", since the effective combination of the Google news API (the crawler) and the text classifier, judging articles for their relevance, is an interaction that determines the logic of what is included in the news-feed and what is rejected.

4.3. Topic and Named Entity Analysis

As articles arrive through the news-feed API, the parsed text from URIs are sent to an Latent Dirichlet Allocation (LDA) topic modeler, and a pre-trained Named Entity Recognition (NER) model is used to identify semantic textual features of each document, as discussed earlier in natural language features section (Section 3). To understand how effectively the natural language features can distinguish between the two document classes, we carried out a statistical analysis on the distributions of LDA topics and NEs in the ENERGY-HUB dataset.

The distribution of topics in each class is formed on the basis of the number of documents belonging to a particular topic, i.e., the inverse document frequency of topics.

A document d is assigned to topic t_i if t_i has the largest probability among the topic set $T = \{t_1, t_2, ..., t_n | n = 100\}$ generated by a topic modeler within document d:

$$topic_d = argmax(i \in \{1..100\} P(t_i|d)). \quad (1)$$

The Named Entity (NE) distribution is formed by counting the number of documents consisting of a particular NE, i.e., the inverse document frequency measure of the NE. The Jensen-Shannon Divergence (JSD), also known as "Information Radius", measures the distance between two probability distributions. In Reference [36], authors used JSD to calculate the distance between two corpora by forming probability distributions of vocabularies using term frequencies. Similarly, the same authors used the JSD method over topics and named-entity distributions over the two sets of documents corresponding to the two classes to compute their divergence.

The JSD measure is calculated using Equation (2) below, where P and Q represent the topic (or NE) distribution in the relevant and irrelevant classes, respectively, and $M = \frac{1}{2}(P + Q)$ and $D(P||Q)$ is an asymmetric distance measurement based on Kullback-Leibler Divergence.

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M). \quad (2)$$

The JSD distance between topic distributions in "relevant" and "irrelevant" classes from the corpus is found to be 0.80, while the JSD distance between the NE distributions is approximately 0.85. These high distance scores provide initial evidence that the natural language features based on topics and NEs can effectively distinguish between the two document classes. Hence, these features are considered as inputs to the neural network classifier which will be described in the Learning Models section below (see Section 5). Triples are also extracted and formed from complete sentences at this time as follows.

4.4. Topic-Entity Triples

The extracted named entities E and LDA topics T from a document D can be used to form triples (linked data), and these often constitute atomic facts. A dependency parser that identifies "subjects" and "objects" in each sentence, while a POS (Part-of-Speech) tagger marks each token in the sentence with an appropriate part-of-speech, as shown in Figure 2. As the text from web pages is mostly semi-structured, namely not always grammatically well-formed, some sentences cannot be rendered as triples. Mapping the extracted named entities against subjects and objects helps resolve this. For any document D, either the subject (or objects) of a sentence must be in its list of entities E, or from the list of keywords from the dominant topics T, to be considered as a candidate triple as defined in Algorithm 1. A well-formed formula in predicate calculus needs a subject, an object and a verb as a predicate. Likewise, a triple must have all three of these elements: namely object, subject, and predicate. These triples are used as an additional feature set as inputs into the training of the NN-classifier to classify relevance/irrelevance documents arriving to the news-feed. The question to answer is, do these triples contribute in a positive way to a relevance/irrelevance classifer?

For instance, consider the two triples formed from the sentence *"U.S. scientists call for fight against climate change"*, namely <U.S. scientists, call, against climate change>, <U.S. scientists, call, for fight> presented earlier. *"U.S. scientists"* is identified as a named entity (by NER), and *"climate change"* is a keyword from the most dominant topic of that document. This triple contains a named entity in its subject and a topic keyword in its object; hence, it can be considered as topic-entity triple. On the other hand, *"fight"* is neither represented in the topic list for the document nor is it a named entity (by NER). Hence, the triple <U.S. scientists, call, for fight> is discarded; it is not a topic-entity triple.

Preliminary analysis (using JSD above in Section 4.3) on the proposed natural language features shows named entities E and LDA-derived topic distributions T have significant information to distinguish the two classes, relevant versus irrelevant documents. Triples

formed from these features are also inferred as a distinguishing feature between two classes, hence being considered as an additional feature for text classification.

Algorithm 1: Triple Extraction Algorithm from Raw Text

Input : Given a document d from the corpus, with abstract topics T and named entities E

Output: A list of Topic-Entity Triples <Subject, Verb, Object>

`TripleExtraction(`d, T, E`)`
I←newlist
for s *in* $d.sentences$ **do**
  $\tau = ExtractTriple(\mathrm{s})$
  if $\tau.subject$ *in* `filter(`T, E`)` *or* $\tau.object$ *in* `filter(`T, E`)` **then**
    I← $addtolist(<\tau.\mathrm{subject}, \tau.\mathrm{verb}, \tau.\mathrm{object}>, \mathrm{I})$
  end
end
return I

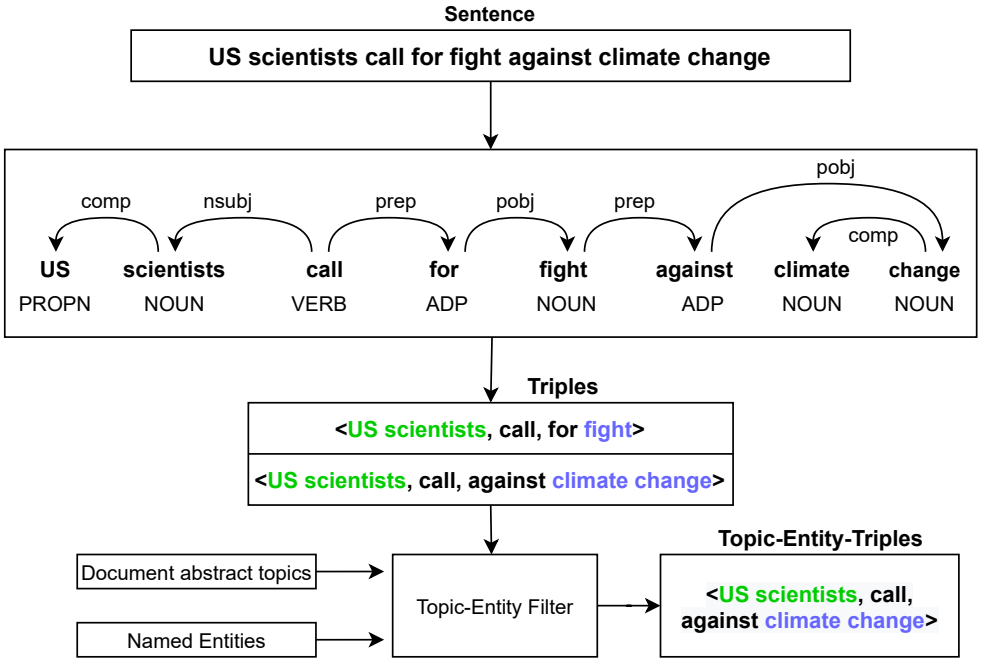

Figure 2. An example of *Topic-Entity Triple* extraction from a sentence in ENERGY-HUB corpus. A pre-trained spaCy model is applied to get syntactic dependencies of each term and subsequently form Triples. Each triple is passed through a filter to compare with extracted named entities, and topic keywords in order to consider them as Topic-Entity Triples.

5. Learning Model Methodology

To discover the influence of semantic and structural text features on news-feed classification effectiveness, we develop three baseline deep learning models: (i) with pre-trained word embeddings as inputs only, (ii) with sentence embeddings as inputs only, and (iii) with bi-directional encoder representation of transformers. Then, probabilistic LDA-based topic distributions, named entities (NEs), and linked data (triples) are added as extra input features to each of the baseline models. Several combinations of features sets are formed to test the performance of the resulting classifiers. Before building these models, common pre-processing steps, like tokenization, special character filtering, and lowercasing, are applied as necessary on the raw text of the news items. The tokens are represented as

strings which are then converted into random unique integers to encode them as vectors. The details of each model is given in the following sections.

GloVe [9] and InferSent [11] were selected for bag-of-word and sentence representations, respectively. As per Reference [11], these representations show superior performance in several NLP tasks compared to other encoders in their class. The BERT model is selected from transformers. BERT and InferSent are considered as state-of-the-art models in text classification problem. Our proposed model architecture enhances the document representations using latent topics, named entities (NEs), and linked data triples (formed from topics and NEs).

5.1. Models with Pre-Trained Word Embeddings

Firstly, a vanilla model is developed with plain text inputs extended using GloVe (Global Vectors for Word Representation). The "input text" section from Figure 3 represents this model's architecture. Given a set of documents from the corpus, each document D is represented as a bag-of-words w_1, w_2, w_3,....w_N. This baseline model is built with a multi-layered feed forward network. The maximum length of the document (N) is fixed as 1000 tokens and given as input to the embedding layer. Vector representations for each token are extracted from pre-computed 300-dimensional Global Vectors (GloVe). The output from the embedding layer is flattened to form a 1000 $\times$ 300 sized vector. This is then passed into a series of fully-connected Rectified Linear Unit ("ReLU") layers of 512 and 100 neurons, respectively. The output "Softmax" layer provides the probability of pre-defined (relevant/irrelevant) classes.

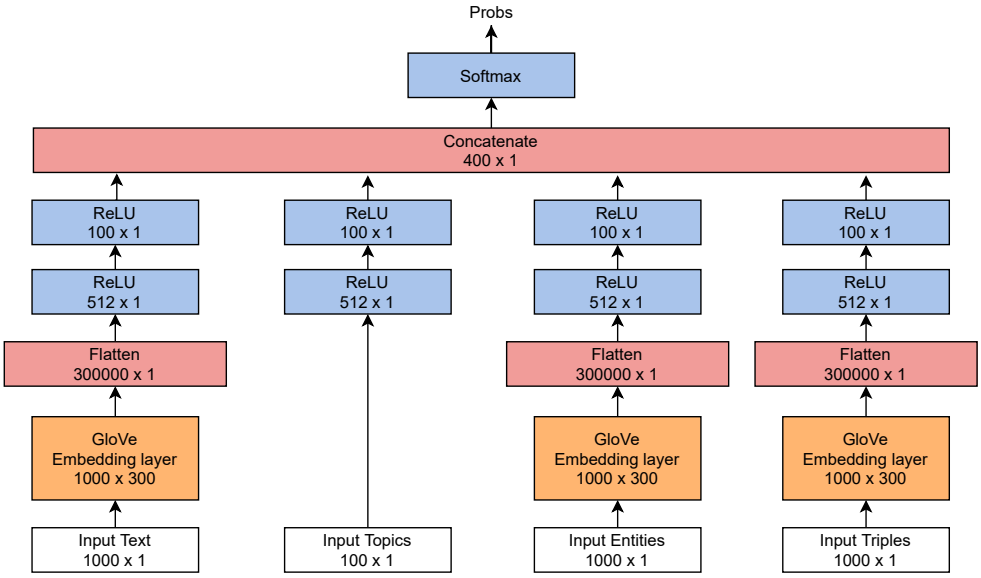

Figure 3. A deep learning architecture for the classification of news-feeds using raw text, topics, named entities, and triples.

Multiple models are then created by the addition of semantic rich elements, LDA topics, named entities, and triples, and these elements are concatenated to the baseline "input text" model. The final models use the combination of all inputs—plain text, LDA-generated topics, named entities, and extracted topic-entity triples.

Input topics are represented as the probability distribution of topics for each document. The number of topics is a hyper-parameter of the LDA topic modeler, identified using an optimum coherence score (further explained in Section 6.1 below).

The length of named entities and triples to input is set to 1000 tokens to be consistent with the baseline "input text" model.

The classifier requires three separate embedding layers to form 300-dimensional vector representations of the text, NEs, and linked data, as shown in Figure 3. These

representations are flattened to obtain a single dimension vector. This vector is fed to multiple "ReLU" layers (512 and 100 neurons). This produces four 100-unit feature vectors from four different inputs. All these vectors are joined together to form a 400 × 1 feature vector. This is used for classification by connecting to a fully-connected "softmax" layer with output neurons. These neurons represent the predicted probability of the classes.

5.2. Models with Sentence Embeddings

Phrases and sentences often provide better contextual meaning to a document than individual words [37]. Sentence representations can be used in place of GloVe word-embeddings to capture word and phrase relationships more effectively. A pre-trained sentence embedding, InferSent [11], encodes each sentence from raw text into a 4096 dimension vector. A document comprises of multiple sentences, S_1, S_2, S_3,...S_n, where S_i is a vector representation of a sentence, and n denotes number of sentences in that document. These sentence representations form a document and undergo entry-wise summation, forming a 4096-sized feature vector. The "sentences" component from Figure 4 is considered as the baseline model. Two "ReLU" layers with 512 and 100 neurons, respectively, form a feed-forward neural network, followed by an output "Softmax" layer, to obtain the predicted probability of the classes.

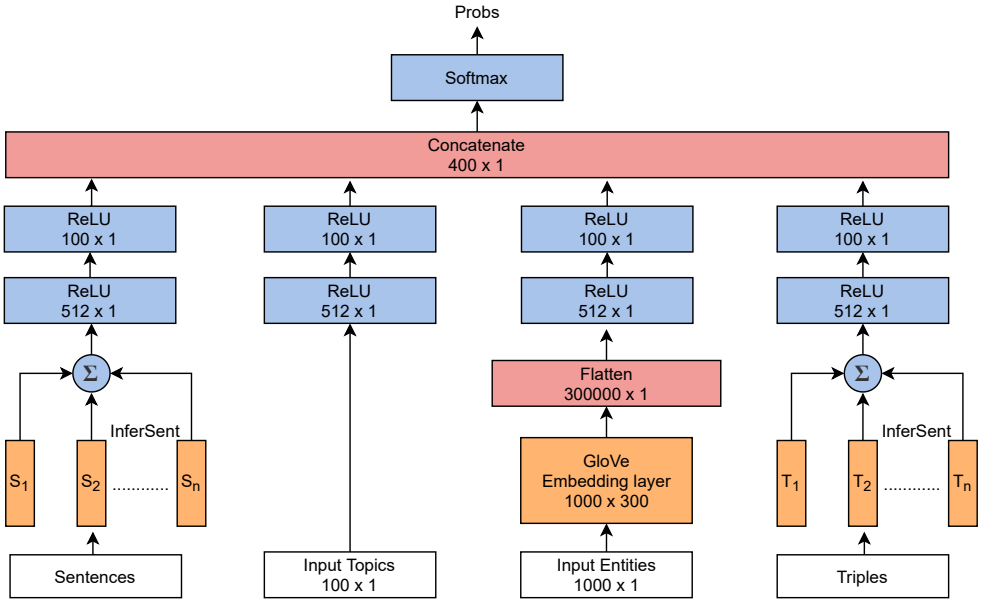

Figure 4. Deep learning architecture for classification of news-feeds using sentence embeddings of text and triples, topics, and named entities.

As document consists of multiple sentences, and a sentence may consist of multiple triples. The sentences formed from these triples are encoded into 4096-vector representation and undergo entry-wise summation similar to the text model. The final model from using sentence embedding is a combination of LDA-derived topics, named entities, and triples formed by them. The textual and triples components of the word embeddings model are replaced with sentence embeddings, while topics and named entities are still represented with topic distributions and GloVe embeddings, as shown in Figure 4. Each input produces a 100-length vector which is concatenated to form a 400 × 1 feature vector. A "softmax" activation is applied on this vector with output neurons for classification.

5.3. Bi-Directional Transformer Based Models

Pre-trained language models, like BERT, have made a significant impact on several natural language tasks, including text classification. The BERT base model contains 12 blocks of transformers, with each having 12 self-attention heads. BERT accepts a maximum

of 512 tokens per document as its input. The model has a hidden size of 768 which encodes the document into a 512 × 768-dimensional vector, as shown in Figure 5. Unlike a bag-of-words model as in Section 5.1, BERT provides self-attentive, contextualized word representations based on neighbor tokens. This sequence output is averaged and fed to a 100-neuron "ReLU" layer, followed by a fully connected "softmax" layer, to predict class probabilities.

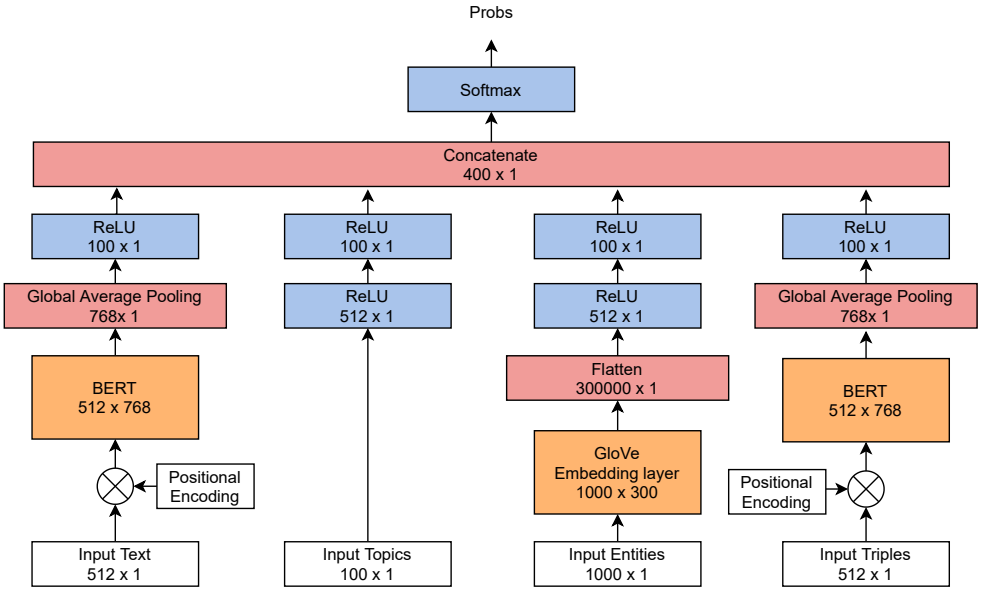

Figure 5. Deep learning architecture for classification of news-feeds with BERT using text, topics, named entities, and triples.

Similar to previous models, additional semantic features are given as inputs to the text-based BERT. Text and triples are applied over separate BERT models. A 100-length feature vector is produced by each model which is concatenated and connected to the output "softmax" layer.

6. Experimental Description

6.1. Extracting Natural Language Features

Latent Dirichlet Allocation (LDA) [38] is adopted for mining topics as semantic features in the ENERGY-HUB. LDA determines clusters of documents over a given number of topics. Gensim (https://radimrehurek.com/gensim/, accessed on 05 February 2021) is a popular software library in Python that implements LDA. It efficiently manages memory allocation when dealing with large corpora [39], hence being a natural choice for our purpose.

After collecting articles through a crawler, these are pre-processed by converting the documents to sentences. Special characters are avoided as they are of little use in topic modeling. Stop words are removed, and bi-grams are generated from the textual data. Lemmatization is performed on the processed text to consider different words with similar meaning as single items. LDA is then applied on the processed corpus to generate topic models.

Identifying the optimum number of topics required helps with better segmentation. A coherence score measures the performance of a topic model by computing the semantic similarity between dominant words in each topic [40]. ENERGY-HUB iterates over 10 to 100 topics, calculates the coherence score for each model, and selects the optimal model with the highest coherence score.

For Named Entity Recognition (NER), we used the Python-based library spaCy [41] on the document text. The library builds on top of a Convolutional Neural Network (CNN) using annotated elements and raw text. The performance of NER systems is evaluated

by comparing the extracted NEs with human annotations [32]. spaCy achieved 86% accuracy (https://spacy.io/usage/facts-figures, accessed on 05 February 2021) for NER on the OntoNotes 5 corpus (https://catalog.ldc.upenn.edu/LDC2013T19, accessed on 05 February 2021), hence being considered suitable for our experiments. Named entities extracted for this work include persons, organizations, nations, events, locations, dates, times, numeric quantities, ordinal, and cardinal numerics.

For the extraction of triples, document sentences are annotated with "subject, predicate and object" using Open Information Extraction (OpenIE) from Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/openie.html, accessed on 8 March 2021) library. A topic-entity filter is applied over each document to identify semantically enriched triples (as discussed in Section 4.4).

6.2. ENERGY-HUB

The statistics detailing the three datasets are provided in Table 1. The ENERGY-HUB corpus is stored in a document database and exported to a JSON file for experimentation. Producing the classifier involves several pre-processing steps. The data is partitioned 80% to 20% for holdout validation, where the larger 80% portion goes for training the neural network, while the smaller 20% is the test set, used for testing the classifier to measure its performance. The training set has 1740 samples and is used for model learning in all epochs. Validation data contains 436 records and is used to measure accuracy and loss after each epoch, helping to prevent over-fitting through early stopping.

The test set has 544 elements and is used for model evaluation. All partitions are performed through stratified sampling to ensure a balanced distribution of classes in the training, validation, and testing data partitions.

Table 1. Features of the three datasets used in these experiments. ENERGY-HUB is an in-house dataset, while REUTERS and 20 NEWS GROUPS are widely used standard news classification datasets.

Dataset	Training Samples #	Test Samples #	Average Length	#Classes
ENERGY HUB	2176	544	1574	2
REUTERS	1828	457	140	2
20 NEWS GROUPS	11,314	7532	284	20

6.2.1. Implementation

The LDA topic model is generated from the training data. The probability topics distributions of each document in the training set, and important keywords of each topic are identified. This pre-trained topic model is then applied on the validation and test sets. The named entities are parsed through spaCy, and the Topic-Entity triples are constructed (as described in Figure 2).

Raw text from each document is split into list of sentences with '.' delimiter. These sentences are encoded using a pre-trained InferSent model, with a batch size of 64, 300-dimensional word embeddings, and 2048-dimensional Long-Short Term Memory (LSTM) encoding. Each sentence is represented as a 4096-dimensional embedded vector.

Each entry from training, validation, and test sets contains document text, the LDA generated probability distribution of topics from the document, the named entities, the Topic-Entity triples, and vector representation of sentences. As the ENERGY-HUB data is biased with more "relevant" examples, down-sampling the collection to match the number of "irrelevant" articles makes the learning set more balanced, so additional irrelevant news articles are sourced and added to the training set. This balancing results in a training set with more or less equal numbers of relevant/irrelevant articles to present to the learning algorithm. The balanced training set has 1382 documents, where 691 (exactly half) are relevant, and the remainder are irrelevant.

Text attribute requires additional pre-processing, like stop word removal, filtering special characters, and case conversion. Stop-words are identified from the natural language tool kit (NLTK (https://www.nltk.org/, accessed on 10 March 2021)) and removed from the

text. Special characters and digits are filtered and the text folded to lower case to maintain consistency. Topics, entities, and triples are composed as a list of elements, which are later converted to text, like sentences.

All models were built using the Keras (https://keras.io/, , accessed on 10 March 2021) API in Tensorflow (https://www.tensorflow.org/guide/keras, accessed on 10 March 2021). The experiments were conducted in Google Colaboratory (https://colab.research.google.com/, accessed on 10 March 2021) with 12,288 Gigabytes (GB) of RAM. The TensorFlow models were built on a Tesla P100 GPU with 16,280 MB memory, CUDA version 10.1, and driver version 418.67. The Adam optimizer with a learning rate of $2 \times e^{-5}$ was found to be the best fit for the data. Sparse categorical cross-entropy is considered as the loss function with accuracy as the optimization metric. The same set of hyper-parameters are used for all models.

GloVe and sentence embedding models are trained for 300 epochs with a batch size of 32 and early stopping using generated training and validation sets. BERT models are trained for 3 epochs with a batch size of 4.

6.2.2. Results

We used accuracy (the number of observations correctly made), precision (the percentage of relevant documents in those retrieved), recall (the percentage of documents in the retrieval set compared to number known to be relevant), F1 score (harmonic mean between precision and recall), and AUC (Area Under Curve) as evaluation metrics reported in Table 2 and visualised in Figure 6. When comparing models, accuracy is not the ideal metric due to the risk of model bias towards a specific class, hence F1-scores being computed.

Table 2. News-feed classification results on the ENERGY-HUB dataset. †s indicate statistically significant differences (with McNemar's test) with reference to the baseline model. For instance, the addition of triples improves the baseline mode (Text (with Glove)) from 64.5% to 68.9%, a difference that is statistically significant. Note: SEs=Sentence embeddings, NEs=Named entities.

Feature Set	Accuracy	Precision	Recall	F1 Score	AUC
Text (with GloVe)	64.5	82.3	49.8	62.1	76.0
+Topics	64.0	77.9	53.3	63.3	72.7
+NEs	66.2	71.0	71.0	71.0	73.9
+Triples †	**68.9**	**77.8**	65.3	71.0	**76.9**
+Topics+NEs	66.5	71.3	71.3	71.3	74.2
+Topics+Triples	68.0	72.7	72.2	72.5	74.8
+NEs+Triples	60.5	76.0	47.0	58.1	69.1
+Topics+NEs+Triples	67.6	67.0	**87.7**	**76.0**	73.9
Text (SEs)	67.8	76.1	65.3	70.3	76.1
+Topics †	70.4	**81.2**	64.0	71.6	78.4
+NEs	63.8	71.3	63.4	67.1	69.2
+Triples (SEs)	70.6	79.8	66.2	72.4	77.8
+Topics+NEs	65.1	70.3	69.4	69.8	70.2
+Topics+Triples (SEs) †	**73.2**	78.4	**74.4**	**76.4**	**79.3**
+NEs+Triples (SEs)	62.3	69.7	62.5	65.9	68.2
+Topics+NEs+Triples (SEs)	63.6	68.1	70.7	69.3	68.4
Text (BERT)	71.3	**88.5**	58.3	70.3	84.3
+Topics †	75.1	85.8	68.7	76.3	86.1
+NEs†	76.4	85.6	71.6	78.0	**86.8**
+Triples (BERT) †	**77.7**	78.0	**86.1**	**81.8**	86.1
+Topics+NEs †	75.9	81.8	75.3	78.4	85.2
+Topics+Triples (BERT) †	76.6	88.3	69.0	77.5	86.6
+NEs+Triples (BERT) †	77.5	82.1	78.5	80.3	84.6
+Topics+NEs+Triples(BERT) †	75.5	81.2	75.3	78.2	85.4

Consider the second block of results in Table 2. These show the performance of the additional features compared to a baseline model consisting of Text (with sentence embeddings), shown in the table as Text (SEs). The sentence embeddings-based model performs better than GloVe-based models with sentence vector representations of text achieving 67.8% accuracy and 70.3% F1 score. The "Text (SEs)+topics", "Text (SEs)+triples", and "Text (SEs)+topics+triples" models perform better than the baseline "Text (SEs)" model. Comparing all sentence embedding-based models, "Text+topics+triples" resulted in the best "F1" of 76.4%.

The third block of results in Table 2 uses BERT as the baseline model for presenting text. The BERT-based models show even better performance than GloVe and sentence embeddings. Considering "Text (BERT)" as a baseline, adding topics, named entities, and triples has positive impact by improving classification performance. Overall, BERT-based "Text+triples" proved to be the best model with an F1 score of 81.8%, and it performed significantly better than the baseline "Text (BERT)" model.

The performance of a neural network is affected by initialization of weights [42]. The variation in performance of a few proposed models is very small compared to the baselines. To make reliable conclusions, the classification results of each model at document level is compared to its baselines. McNemar's test helps to identify the difference between two predictions by calculating the number of wrongly classified documents by each model. The models in the table tagged as statistically significant have a p-value less than 0.05 obtained from the McNemar's test.

The "Text (with GloVe)+Triples", "Text (SEs)+topics+triples", and "Text (BERT)+triples" models pass the significance test compared to their respective baseline models. This demonstrates that the addition of our proposed semantic and structural features results in improved news-feed classification even in combination with a complex sentence embeddings or state-of-the art transformer models as a baseline. Figure 6 provides a visualisation of results from Table 2. The latent token-based information in topics, as well as the surface NEs and semantic relationships encapsulated in triples, thus, are shown to complement the deep semantics captured by the pre-trained sentence and word embedding information when scoring relevance/irrelevance of news-feed items.

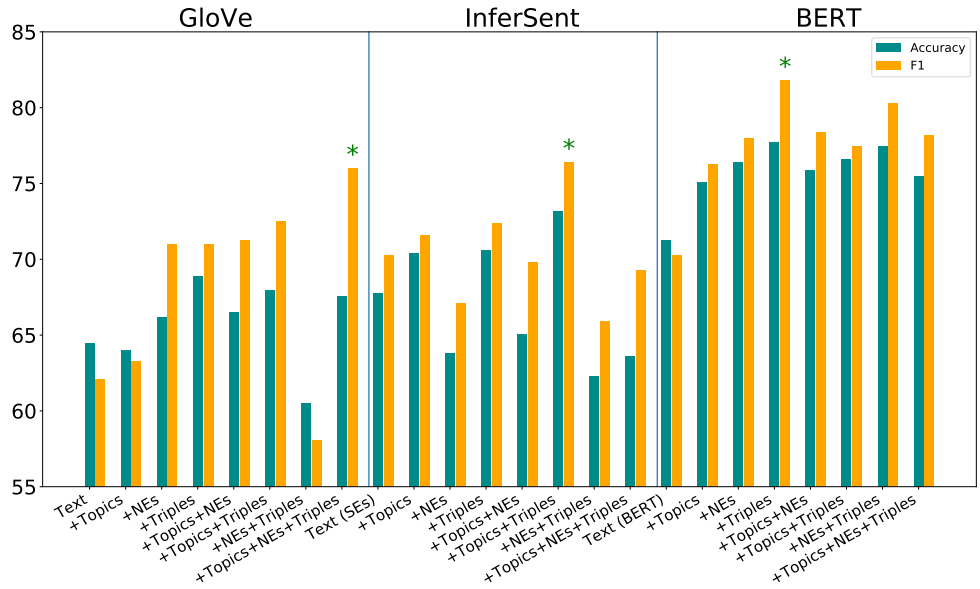

Figure 6. Performance results of classifiers from different inputs on ENERGY-HUB dataset. The first block of results uses GloVe-based representations. The second is models using the InferSent sentence encoder. The third block of results uses BERT-based models. ∗ indicate the best model from each class

6.3. REUTERS *and* 20 NEWS GROUPS

To test the generality of the results obtained from the ENERGY-HUB dataset, the same experiments were repeated on two standardized datasets, namely REUTERS sourced from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection, accessed on 5 February 2021) and 20 NEWS GROUPS. The REUTERS corpus is a multi-class, multi-labeled dataset. It has 21,578 instances categorized into 90 classes. ModApte version of split data with 9603 training documents and 3299 testing documents of 10 largest classes is used in the experiments. The 20 NEWS GROUPS is a multi-class news-topic categorization dataset with 20 different classes. It has 11,314 training samples and 7532 testing samples.

To maintain consistency with ENERGY-HUB corpus, one of the two standard datasets, namely REUTERS, is restricted to two class labels. All documents consisting of particular label ("earn"), a label class with the dataset that allows it to simply convert to a binary class dataset, were considered "relevant", and the remainder were categorized as "irrelevant". Two thousand two hundred and eighty-four documents were sampled where the relevant and the irrelevant set contained 1142 records. The training set with 1462 documents was used for learning. The validation and test sets contained 366 and 456 documents, respectively. The training samples from 20 NEWS GROUPS dataset were partitioned into two sets, forming 9051 training and 2263 validation instances to measure loss after each epoch. All models were trained with the same neural network architecture and parameters used in the ENERGY-HUB tests. Table 3 and Figure 7 shows the results on the REUTERS dataset.

Table 3. News-feed classification results on the REUTERS dataset. †s indicate statistically significant differences (using McNemar's test) with reference to the baseline models. Note: SEs=Sentence embeddings, NEs=Named entities.

Feature set	Accuracy	Precision	Recall	F1 Score	AUC
Text (with GloVe)	86.4	87.8	78.6	83.0	93.4
+Topics	86.4	88.2	78.1	82.9	93.0
+NEs	85.8	88.5	76.0	81.8	91.5
+Triples†	**89.9**	**90.1**	85.4	**87.7**	**95.0**
+Topics+NEs	86.0	80.8	**87.5**	84.0	94.4
+Topics+Triples	87.7	90.0	79.7	84.5	93.4
+NEs+Triples	86.0	82.0	85.4	83.7	93.8
+Topics+NEs+Triples	88.2	85.9	85.9	85.9	94.1
Text (SEs)	95.8	94.8	95.3	95.1	99.4
+Topics	**96.3**	93.5	**97.9**	**95.7**	**99.5**
+NEs †	89.1	86.6	87.5	87.0	94.5
+Triples (SEs)	**96.3**	**96.3**	94.8	95.5	99.3
+Topics+NEs †	91.0	90.8	87.5	89.1	95.8
+Topics+Triples (SEs)	**96.3**	94.9	96.4	95.6	99.3
+NEs+Triples (SEs) †	90.6	89.8	87.5	88.7	96.0
+Topics+NEs+Triples (SEs) †	91.5	92.7	86.5	89.5	95.9
Text (BERT)	98.4	96.9	**99.4**	98.2	**99.8**
+Topics	98.2	97.9	97.9	97.9	**99.8**
+NEs †	95.6	90.9	**99.4**	95.0	99.6
+Triples (BERT)	97.5	**99.4**	94.7	97.0	99.7
+Topics+NEs	97.3	96.8	96.8	96.8	99.7
+Topics+Triples (BERT)	98.0	98.9	96.3	97.6	**99.8**
+NEs+Triples	**98.9**	**99.4**	97.9	**98.6**	**99.8**
+Topics+NEs+Triples (BERT)	98.0	98.4	96.8	97.6	99.6

Consider the first block of results in Table 3, which shows the performance of the additional features compared to a baseline model consisting of Text (with GloVe) word embeddings. "Text (with GloVe)+Triples" achieved the best results with 89.9% accuracy

and 87.7% F1 score. The "text+topics+triples" and "Text (with GloVe)+topics+NEs+triples" models showed better performances than the baseline text model with respect to all of the evaluation metrics. These results are, therefore, consistent with those from ENERGY-HUB tests, namely reinforcing that adding triples improves text classification performances.

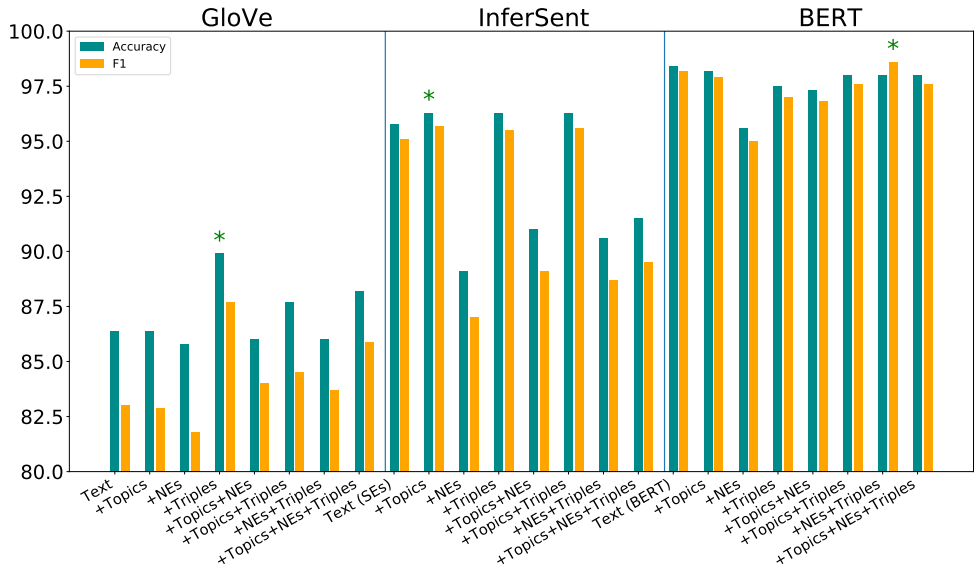

Figure 7. Performance results of classifiers from different inputs using the REUTERS dataset. The first block of results uses GloVe-based representations. The second block is from models using the InferSent sentence encoder. The third block uses BERT-based models. ∗ indicate the best model from each class

Consider the second block of results in Table 3, which shows the performance of the additional features compared to a baseline model consisting of Text (with sentence embeddings), shown in the table as Text (SEs). A similar pattern of results are observed consistent with those achieved with the ENERGY-HUB dataset. The addition of our proposed lexical semantic and structural features to the baseline "Text (SEs)" model produced better results, although the model results are not significant according to McNemar's test. "Text (SEs)+Topics"', "Text (SEs)+Triples", and "Text (SEs)+Topics+Triples" show marginal improvements over the baseline "Text (SEs)" model.

BERT models are the best performing ones with each model producing less than 10 wrongly classified documents. "Text (BERT) + NEs + Triples" is the best performing model with an F1 score of 98.6% compared to 98.2% over the baseline BERT.

The 20 NEWS GROUPS is a multi-class dataset, so metrics, like Macro F-score (average F-score for each class) and Weighted F-score (average F-score considering class proportions), are used in addition to accuracy. Similar to the earlier experiments, adding proposed features to the plain text (baseline) ("Text (GloVe)") increased the classification metrics, with the best being "Text (GloVe)+Topics+NEs+Triples" in Table 4. On sentence embedding and BERT models, "Text+Topics" achieved the best performance, while triples delivered only slight improvements to the baseline models. Abstract LDA topics are an important feature set for classifying 20 NEWS GROUPS and have shown significant improvement in Figure 8, this being the precise way they are organized.

To conclude, adding semantic feature-based models significantly improves performance when compared to baseline "text" models, although the additional features, once added, do not have a significant impact on the baseline sentence embedding or BERT models. The important aspect of the experiments on the REUTERS and 20 NEWS GROUPS datasets is the consistency of the results with those achieved on the ENERGY-HUB dataset. The repeatability of the results for these generic datasets validates our findings as gen-

eralizeable to a broader domain of text classification than news-feed items used in the ENERGY-HUB.

Table 4. News-feed classification results on the 20 NEWSGROUPS dataset. †s indicate statistically significant differences (using McNemar's test) with reference to the baseline models. Note: SEs=Sentence embeddings, NEs=Named entities.

Feature set	Accuracy	Macro-F1	Weighted-F1
Text(GloVe)	49.7	48.5	49.3
+Topics †	52.1	50.8	51.8
+NEs †	54.2	52.9	53.6
+Triples	50.4	49.7	50.6
+Topics+NEs †	55.0	54.3	55.2
+Topics+Triples †	52.5	51.4	52.4
+NEs+Triples †	54.7	53.2	54.1
+Topics+NEs+Triples †	**56.1**	**55.0**	**55.9**
Text (SEs)	74.1	73.4	74.2
+Topics †	**74.8**	**73.9**	**74.8**
+NEs †	59.5	58.9	59.7
+Triples (SEs)	74.1	72.9	73.9
+Topics+NEs †	62.2	61.3	62.1
+Topics+Triples (SEs) †	74.6	73.9	74.8
+NEs+Triples (SEs) †	59.1	58.8	59.5
+Topics+NEs+Triples (SEs) †	61.8	60.7	61.6
Text (BERT)	83.3	82.6	83.3
+Topics †	**84.3**	**83.6**	**84.4**
+NEs †	82.4	81.5	82.3
+Triples (BERT)	83.9	83.1	84.0
+Topics+NEs	83.8	83.1	83.8
+Topics+Triples (BERT) †	82.2	81.5	82.3
+NEs+Triples (BERT)	83.3	82.3	83.1
+Topics+NEs+Triples (BERT)	83.5	82.4	83.3

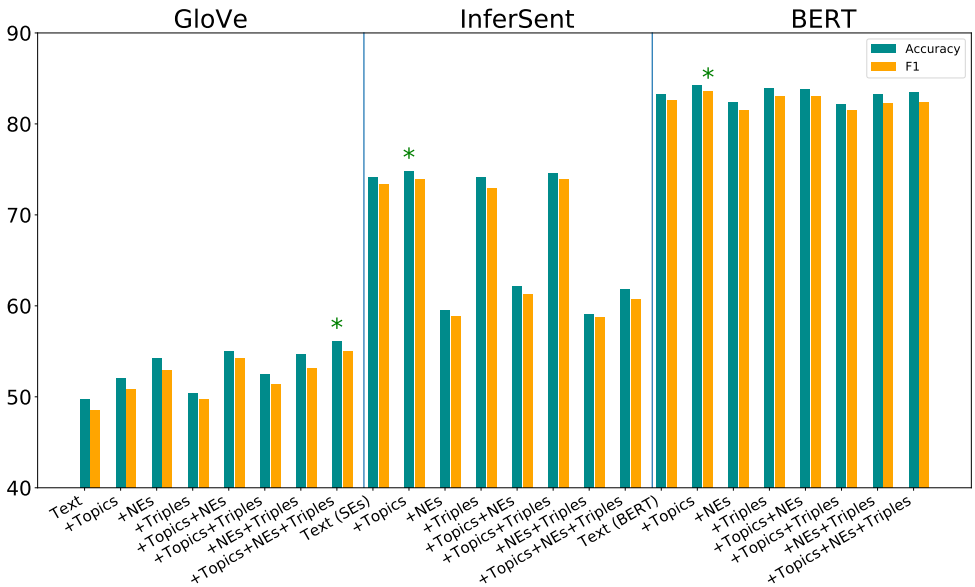

Figure 8. Performance results of classifiers from different inputs on the 20 NEWSGROUPS dataset. The first block of results uses GloVe-based representations. The second block is models using the InferSent sentence encoder. The third block uses BERT-based models. ∗ indicate the best model from each class

6.4. Effect of Document Length on BERT Models

BERT outperforms other models by representing contextualized and self-attentive word vectors. The pre-training of BERT teaches the model to understand language structure by learning linguistic features through its projections. A study by Reference [43] finds that each attention head in BERT attends to a different set of features. Though the features are not explicitly defined, the attention heads of BERT extract syntactic relations of subject-verb-object. This makes the triple structure redundant when added to the baseline text model and accounts for the absence of any significant performance gain.

BERT also restricts its input vector length to 512 tokens, so important information from long documents is truncated. This impacts model performance. An interesting finding from our experiments is that combining "Triples" with "Text" input results in significant performance improvement on long documents. This is because the text model of BERT is restricted to the first 512 tokens, so adding Triples enriches the model beyond the plain text. Triples extracted from document sentences are mostly shorter than plain text and, therefore, provide a valuable abbreviated input to the BERT model.

To validate this claim, long documents (those with more than 600 words) are considered from ENERGY-HUB and the 20 NEWS GROUPS dataset. The REUTERS dataset has very few long documents, thus being unsuited for this comparison.

The results from Table 5 show that classification performance on long documents is improved for the ENERGY HUB and standard 20 NEWS GROUPS datasets. McNemar's tests are conducted on these models, and it is found that long documents produce significantly different results. This provides further evidence to our claim that triples constructed using topics and named entities contain significant latent semantic information and can be used to enrich document representations.

Table 5. BERT model results on long documents from ENERGY-HUB and 20 NEWS GROUPS datasets. Note: Long documents are filtered from the larger corpus using word length (documents with less than 600 words are excluded).

Dataset	Document Type	Feature Set	Accuracy	F-Score
ENERGY-HUB	All	Text	71.3	70.3
ENERGY-HUB	All	Text + Triples	77.7	81.8
ENERGY-HUB	Long documents	Text	71.0	68.9
ENERGY-HUB	Long documents	Text + Triples	74.4	75.4
20 NEWS GROUPS	All	Text	83.3	82.6
20 NEWS GROUPS	All	Text + Triples	83.9	83.1
20 NEWS GROUPS	Long documents	Text	71.9	67.7
20 NEWS GROUPS	Long documents	Text + Triples	75.7	71.3

7. Discussion

Latent topics generated by topic modeler provide global semantic meaning [44] for the words in a document. Addition of topic probability distributions enriched the word/sentence representations and improved their classification performance. As a bag-of-words technique, named entities have shown positive effect on GloVe and Transformer models, but they negatively impact the sentence embeddings. Our proposed method considers if linked data triples, containing lexical relationships along with terms from LDA topics and named entities, contain enriching semantic information. When LDA topics and named entities are fused with generalized word/sentence embeddings, as in Section 5, an enriched document representation is created. The addition of these features results in improved performance for a relevance classifier, compared to the baseline models which do not contain them. In contrast with other works, our proposed methodology shows performance improvements without recourse to external knowledge sources, dictionary, ontology, or other knowledge sources.

Our finding shows that topic-entity triples contain significant semantic value. Triples created from sentences, filtered and formulated using LDA topics and named entities, can

be used to create information retrieval systems, such as focused search engines, and we posit that they can be applied in question answering since they represent raw facts about a document's meaning.

In practice, our proposed method has some limitations. As topics evolve, the performance of the model may decrease over time [45], and this called "topic drift". One way to address topic drift is to re-train the topic modeler with an updated corpus at regular intervals. Multi-modal architectures provide the ability to merge different models (created for each input) to perform a single task. Creating multiple models increases the number of trainable parameters and often requires greater computing resources and longer training times. The issue of document length, and its impact on classification using BERT by virtue of the limitation of 512 input tokens, is an issue in its own right, and this issue is comprehensively dealt with elsewhere, in Reference [46,47].

8. Concluding Remarks

In this work, we developed the document relevance NN-classifier with three independent models applied on text, topics, and named entities (NEs), and these features are merged to form as a single multi-model classifier, leading to a higher-fidelity predictor with increased trainable parameters, called the *Topic-Entity Model*. The key question became: could the inclusion of topic-entity triples improve performance or convergence speed of a relevance classifier? We show that it can. The importance of this finding is that it demonstrates that triples add latent semantic enrichment to the relevance/irrelevance NN-classifier. We call this the *Topic-Entity-Triple Model*.

The results in all datasets indicate that adding our proposed semantic features improves the performance of the classifier when compared with the Glove-based bag-of-words model. The F-score of the semantic feature-enriched models on the Glove representations is improved over the baseline by 6%. GloVe representations on ENERGY HUB is on par with that of the computationally expensive and state-of-the-art sentence embeddings-based classifier. The performance of the state-of-the art InferSent and BERT encoders also show significant improvements using our proposed models over their respective baselines.

Furthermore, the ENERGY-HUB dataset consists of news articles scraped from real-world web pages with huge amounts of noise in the form of advertising content, social media plugins, and other article recommendations, unlike REUTERS, which is a clean benchmark dataset. Our observations, therefore, reveal that the proposed semantic features extracted from the raw, noisy text play a more significant role in separating the two classes of documents as compared with situations where there is noise-free text in documents.

This paper proposes a novel way of testing the latent semantic value of triples (linked data) by experimenting with state-of-the-art NLP-techniques. Document text, LDA topics, and named entities act as core elements in the formation of triples that are added as features to train a deep neural network relevance classifier. We showed that the best performing classifier for document relevance is one that is formed by a complex network architecture with plain text and added semantic features that are constructed by connecting LDA topics, named entities, and topic-entity triples. This validates the hypothesis that topics, named entities, and triples (linked data) constructed from these NLP-elements enrich semantic features in determining document relevance/irrelevance, especially in the context of news-feed classification.

Author Contributions: Conceptualisation, D.N. and P.W.E.; methodology, D.N., P.W.E., B.O. and M.R.B.; software, D.N.; validation, P.W.E., B.O. and M.R.B.; writing–original draft preparation, D.N., P.W.E., writing–review and editing, P.W.E., B.O., M.R.B.; supervision, P.W.E.;

Funding: This research is funded by "Minerals Council of Australia".

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not Applicable

Data Availability Statement: Data is included within the article

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Paek, T.; Gamon, M.; Counts, S.; Chickering, D.; Dhesi, A. *Predicting the Importance of Newsfeed Posts and Social Network Friends*; AAAI; Fox, M., Poole, D., Eds.; AAAI Press, Atlanta, Georgia, USA: 2010.
2. Setty, S.; Jadi, R.; Shaikh, S.; Mattikalli, C.; Mudenagudi, U. Classification of facebook news feeds and sentiment analysis. In Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Delhi, India, 24-27 September 2014; pp. 18–23.
3. Bhosale, G.; Jain, K.; Aboli Kalyankar, A.; Patange, S. Social Networking News Feeds Classification Using Naïve Bayes and Support Vector Machine and Watchdog. *Int. J. Innov. Res. Comput. Commun. Eng.* **2017**, 7550–7555, doi:10.15680/IJIRCCE.2017.0504186.
4. Barberá, P.; Boydstun, A.; Linn, S.; McMahon, R.; Nagler, J. Automated Text Classification of News Articles: A Practical Guide. *Political Anal.* **2020**, 1–24, doi:10.1017/pan.2020.8.
5. NLP in News Feeds. Available online: https://syncedreview.com/2019/01/12/nlp-in-news-feeds/ (accessed on 26 September 2020).
6. Blei, D.; Ng, Y.; Jordan, M. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
7. Kim, H.; Sun, Y.; Hockenmaier, J.; Han, J. ETM: Entity Topic Models for Mining Documents Associated with Entities. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, 10–13 December 2012; pp. 349–358.
8. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. *CoRR* **2013**, Arxiv:abs/1310.4546.
9. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25-29 October 2014; pp. 1532–1543.
10. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *CoRR* **2016**, arxiv:abs/1607.04606.
11. Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; Bordes, A. *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data*; EMNLP; Palmer, M., Hwa, R., Riedel, S., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 670–680.
12. Cer, D.; Yang, Y.; Kong, S.y.; Hua, N.; Limtiaco, N.; John, R.S.; Constant, N.; Guajardo-Céspedes, M.; Yuan, S.; Tar, C.; et al. Universal sentence encoder. *arXiv* **2018**, arXiv:1803.11175.
13. Mnih, A.; Hinton, G.E. A Scalable Hierarchical Distributed Language Model. In *Advances in Neural Information Processing Systems*; Koller, D., Schuurmans, D., Bengio, Y., Bottou, L., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2009, Volume 21.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *CoRR* **2017**, arxiv:abs/1706.03762.
15. Munir, K.; Sheraz Anjum, M. The use of ontologies for effective knowledge modelling and information retrieval. *Appl. Comput. Informatics* **2018**, *14*, 116–126.
16. Agarwal, S.; Singhal, A.; Bedi, P. Classification of RSS feed news items using ontology. In Pprceedings of the 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), Kochi, India, 27–29 November 2012; pp. 491–496.
17. Yücesan, M.M.; Dogdu, E. *News Clustering Using Linked Data Resources and Their Relationships*.
18. Kastrati, Z.; Imran, A.S.; Yayilgan, S.Y. The impact of deep learning on document classification using semantically rich representations. *Inf. Process. Manag.* **2019**, *56*, 1618–1632.
19. Fellbaum, C. (Ed.) *WordNet: An Electronic Lexical Database*; Language, Speech, and Communication; MIT Press: Cambridge, MA, USA, 1998.
20. Chen, J.; Hu, Y.; Liu, J.; Xiao, Y.; Jiang, H. Deep short text classification with knowledge powered attention. In Proceedings of the AAAI Conference on Artificial Intelligence, Honulu, Hawaii, USA, 27 January - 1 February 2019; Volume 33, pp. 6252–6259.
21. Suchanek, F.M.; Kasneci, G.; Weikum, G. Yago: A large ontology from wikipedia and wordnet. *J. Web Semant.* **2008**, *6*, 203–217.
22. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, Canada, 9–12 June 2008; pp. 1247–1250.
23. Wang, S.; Tang, J.; Aggarwal, C.; Liu, H. Linked document embedding for classification. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, Indiana, USA, 24 - 28 October 2016; pp. 115–124.
24. Shanavas, N.; Wang, H.; Lin, Z.; Hawe, G. Ontology-based enriched concept graphs for medical document classification. *Inf. Sci.* **2020**, *525*, 172–181.
25. Pavlinek, M.; Podgorelec, V. Text classification method based on self-training and LDA topic models. *Expert Syst. Appl.* **2017**, *80*, 83–93.
26. Kumar, A.; Srinivasan, K.; Cheng, W.H.; Zomaya, A.Y. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Inf. Process. Manag.* **2020**, *57*, 102141.

27. Jian, F.; Huang, J.X.; Zhao, J.; He, T.; Hu, P. A simple enhancement for ad-hoc information retrieval via topic modelling. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Tuscany, Italy, 17–21 July 2016; pp. 733–736.
28. Zhang, Z.; Miao, D.; Gao, C. Short text classification using latent Dirichlet allocation. *Jisuanji Yingyong/J. Comput. Appl.* **2013**, *33*, 1587–1590.
29. Arora, R.; Ravindran, B. Latent Dirichlet Allocation Based Multi-Document Summarization. In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*; Association for Computing Machinery: New York, NY, USA, 2008; pp. 91–97.
30. Krestel, R.; Fankhauser, P.; Nejdl, W. Latent dirichlet allocation for tag recommendation. In Proceedings of the Third ACM Conference on Recommender Systems, New York, NY, USA, 23–25 October 2009; pp. 61–68.
31. Yadav, V.; Bethard, S. A survey on recent advances in named entity recognition from deep learning models. *arXiv* **2019**, arXiv:1910.11470.
32. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, **2020**. doi: 10.1109/TKDE.2020.2981314.
33. Rusu, D.; Dali, L.; Fortuna, B.; Grobelnik, M.; Mladenic, D. Triplet extraction from sentences. In Proceedings of the 10th International Multiconference Information Society-IS, Ljubljana, Slovenia, 8–12 October 2007; pp. 8–12.
34. De Marneffe, M.C.; Manning, C. The Stanford typed dependencies representation. In Proceedings of the Coling 2008, Workshop on Cross-Framework and Cross-Domain Parser Evaluation, Manchester, UK, August 2008; pp. 1–8.
35. Lin, D. Dependency-based evaluation of MINIPAR. In *Treebanks*; Text, Speech and Language Technology, vol 20. Springer, Dordrecht, 2003; pp. 317–329.
36. Aamer, H.; Ofoghi, B.; Verspoor, K. Syndromic surveillance through measuring lexical shift in emergency department chief complaint texts. Proceedings of the Australasian Language Technology Association Workshop 2016, Melbourne, Australia, 5-7 December 2016, pp. 45–53.
37. Hill, F.; Cho, K.; Korhonen, A. Learning distributed representations of sentences from unlabelled data. *arXiv* **2016**, arXiv:1602.03483.
38. Priss, U. Lattice-based Information Retrieval. *Knowl. Organ.* **2000**, *27*, 132–142.
39. Rehurek, R.; Sojka, P. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, Citeseer, 22 May 2010.
40. Lapata, M.; Barzilay, R. Automatic evaluation of text coherence: Models and representations. In Proceedings of the IJCAI, Scotland, UK, 30 July–5 August 2005; Volume 5, pp. 1085–1090.
41. Honnibal, M.; Montani, I. spacy 2: Natural language understanding with bloom embeddings. In *Convolutional Neural Networks and Incremental Parsing*; To appear, 2017.
42. Koturwar, S.; Merchant, S. Weight initialization of deep neural networks (DNNs) using data statistics. *arXiv* **2017**, arXiv:1710.10570.
43. Clark, K.; Khandelwal, U.; Levy, O.; Manning, C.D. What does bert look at? an analysis of bert's attention. *arXiv* **2019**, arXiv:1906.04341.
44. Dieng, A.B.; Wang, C.; Gao, J.; Paisley, J. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv* **2016**, arXiv:1611.01702.
45. Al-Ghossein, M.; Murena, P.A.; Abdessalem, T.; Barré, A.; Cornuéjols, A. Adaptive collaborative topic modeling for online recommendation. In Proceedings of the 12th ACM Conference on Recommender Systems, Vancouver, BC, Canada, 2–7 October 2018; pp. 338–346.
46. Pappagari, R.; Zelasko, P.; Villalba, J.; Carmiel, Y.; Dehak, N. Hierarchical transformers for long document classification. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 838–844.
47. Adhikari, A.; Ram, A.; Tang, R.; Lin, J. Docbert: Bert for document classification. *arXiv* **2019**, arXiv:1904.08398.