Automated Due Diligence: Building Knowledge Graphs from News

Ilze Amanda Auzina
Vrije Universiteit
Amsterdam, The Netherlands
ilze.amanda.auzina@gmail.com

Jenia Kim
Vrije Universiteit
Amsterdam, The Netherlands
d.e.kim@student.vu.nl

Evert Haasdijk
Deloitte
Amsterdam, The Netherlands

Frank van Harmelen
Vrije Universiteit
Amsterdam, The Netherlands

Piek Vossen
Vrije Universiteit
Amsterdam, The Netherlands

ABSTRACT

Before a company enters a new business relationship it has to perform a background check, known as due diligence. It is commonly carried out by a human expert and involves screening a large amount of unstructured textual information (e.g. news articles), which is extremely labor intensive. We propose to automate this process, which would allow to, firstly, reduce the time needed for article screening, and, secondly, discover new insights about the network the company operates in. The solution includes (a) a classifier that detects articles containing negative events about the company of interest, and (b) a knowledge graph that combines the gained information with structured data sources. We report promising results of the novel approach to utilize semantic frames of the article's predicates as features for the news article classification. Furthermore, we have successfully built a knowledge graph that combines information from different data sources. The proposed automated pipeline introduces a promising novel alternative for the commonly performed due diligence procedure.

KEYWORDS

automated due diligence, knowledge graph, adverse media detection, semantic frames

1 INTRODUCTION

Due diligence is a process of investigating an organization or person before entering a business relationship with them. In some cases, it is a legal obligation; for example, in 2018 the Dutch bank ING was fined $900 million for failing to properly vet the beneficial owners of clients' accounts, allowing these accounts to be used for money laundering.[1] In other cases, it is a voluntary investigation which contributes to more informed decision making and risk mitigation. Due diligence investigation of a company includes, for example, checking who its ultimate beneficial owner is, if it is listed in any sanction list, if it is involved in illegal or unethical practices (e.g. lawsuits, child labor, corruption, environmental issues) and so on. It also includes checking whether the company or key persons in it are directly or indirectly associated with other parties that might be involved in illegal or unsavory activities.

Some of the information relevant for due diligence can be found in structured databases, such as the national Chamber of Commerce. Other relevant information, however, is not documented in a structured way but can be found in the form of free text, e.g. news articles. Analyzing this unstructured textual information and extracting knowledge from it is extremely labor-intensive. Therefore, the scope of this process is quite limited when done manually: a human expert usually performs a web search on the company name and scans the first 10-20 news headlines to see whether any negative events or worrisome connections are mentioned.

We propose a way to partially automate the due diligence process and make it more informative. Figure 1 shows an overview of the proposed solution, which includes (a) processing a dataset of news items about the company of interest (indicated in green) to detect adverse media, i.e. articles that mention negative events related to the company, (b) extracting the entities (people and organizations) mentioned in these articles into a structured graph, and (c) expanding the graph with additional information from structured resources like DBpedia[2] and the Offshore Leaks Database[3].

In comparison to the manual search done by a human expert, our approach can not only process a larger number of news articles, but also generate new knowledge by linking information from structured and unstructured sources. Representing the information in a knowledge graph allows us to show people and organizations that are directly related to the original seed company, and potentially to identify new indirect connections that could not have been found by investigating each source separately. Moreover, suspicious entities (e.g. entities mentioned in the Offshore Leaks) are flagged to the end user and can be investigated further by inputting them into the pipeline and searching for adverse media about them. This way, the graph can be iteratively expanded with additional relations and entities.

The human expert performing the due diligence receives two outputs from our system: (a) the (suspected) adverse media articles, and (b) the extended graph showing the interlinked network of entities related to the company of interest, with the suspicious entities flagged. The first output greatly reduces the human time and effort involved in the due diligence process, since only relevant articles need to be manually examined and analyzed. The second output provides the expert with new insights, which cannot be easily obtained manually, about the network in which the company operates. The proposed solution allows the human experts to shift

[1] https://www.om.nl/publish/pages/58352/feitenrelaas_houston.pdf

[2] DBpedia is a crowd-sourced open knowledge graph containing information created in various Wikimedia projects: https://wiki.dbpedia.org/.

[3] ICIJ Offshore Leaks database contains information about offshore accounts involved in international tax fraud: https://offshoreleaks.icij.org/

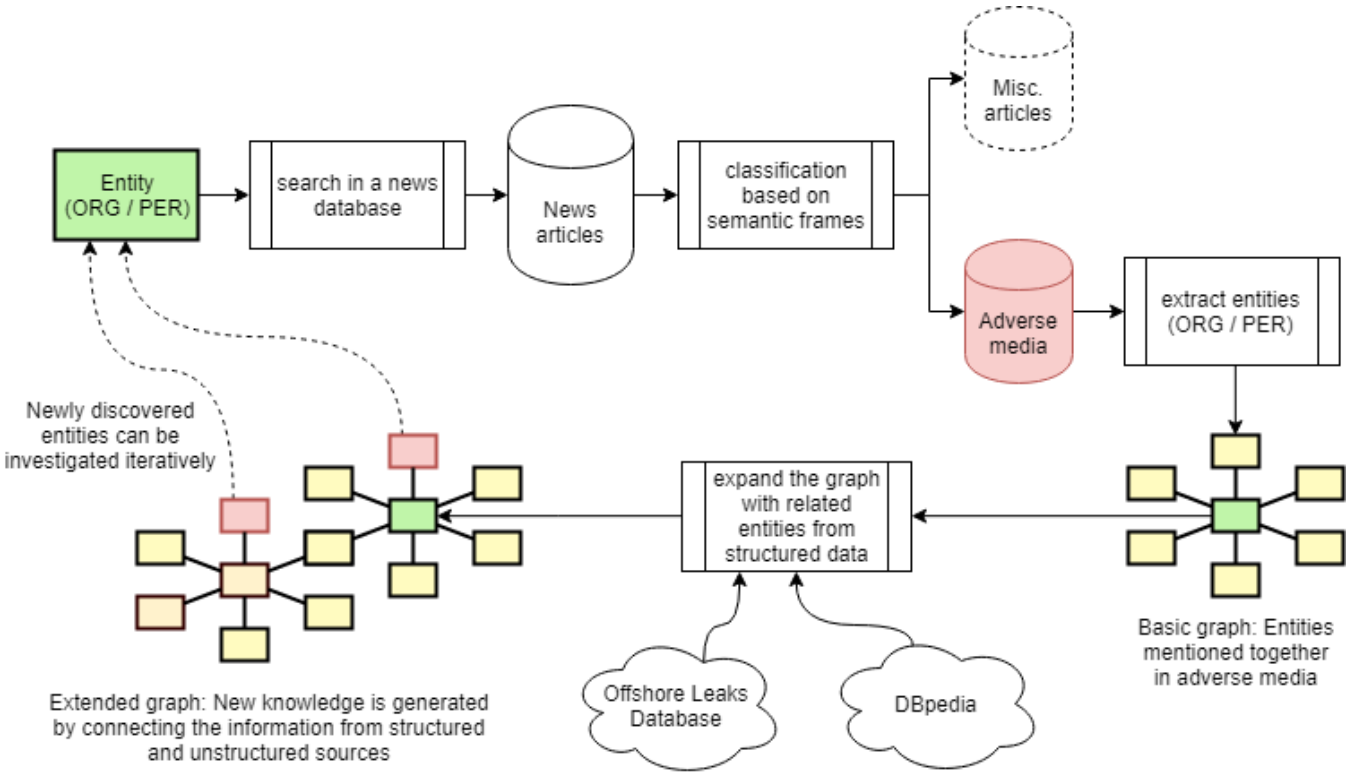

Figure 1: Overview of the proposed automated due diligence solution.

their focus from information extraction and processing to analysis and decision making.

The quality of our system's outputs hinges on our ability to solve two main challenges: (a) successfully filtering out adverse media from a large collection of articles, and (b) merging separate data sources together.

The first task is a text classification problem, whose difficulty lies in the fact that a range of very different topics fall under the umbrella term "adverse media". Accordingly, the words in adverse media articles are very diverse and using them as features for classification creates a very complicated, high-dimensional problem. To address this concern, we use semantic frames as features, instead of words. Semantic frames are a way to generalize beyond specific lexical items to types of events. For example, the frame `Commerce_buy` describes a commercial transaction event; this more abstract representation encompasses various words, such as the verbs *buy* and *purchase*, and the noun *acquisition*. Our hypothesis is that this approach reduces the noise by focusing on events only (i.e. predicates rather than all words), thus generalizing beyond the diverse topics of the articles to capture relevant types of events. The results of the experiment we report in Section 2 provide an indication in favor of this hypothesis. [4]

The second challenge is making links, so that a person or a machine can explore the data. Even though a large amount of structured data is available on the web, in order to query over multiple sources of data at once you have to use a uniform resource identifier (URI) that unambiguously identifies a particular resource. Such an identification allows interaction with representations of the resource over the entire network, hence, knowledge can be obtained from all the resources at once. The Offshore Leaks Database only has an internal unique identifier for every entity, hence in order to merge the information with other resources it is necessary to link the existing identifiers to an external one, which would be present in the other data sources. The entities which were extracted from the news search had a unique DBpedia URI, therefore, we chose to reference all the entities to their corresponding DBpedia page URI. The success and limitations of this approach are discussed in Section 3.2.

The use-case in this paper is provided by the Port of Moerdijk, the 4th sea port of the Netherlands.[5] There are over 400 companies that operate on the port's facilities, and each new company that does business on these facilities (either directly or indirectly through existing customers) needs to be vetted by the port. Our experiment focuses on one of the companies operating in the port: MM Metal Recycling B.V., a subsidiary of the Japanese Mitsubishi Materials Corporation.

In Section 2, we present our method for detecting adverse media in a collection of news articles. In Section 3, we describe how the different data sources are integrated into a knowledge graph. In

[4]Further support could come from a direct comparison of our model with the same algorithm trained on the same data, but with tokens/lemmas as features. We leave this comparison to future work.

[5]https://www.portofmoerdijk.nl/en/

	MM dataset	KS dataset
Source	Nexis Uni[7]	Nexis Uni
Search term	Mitsubishi Materials	Kobe Steel
Content type	news	news
Language	English	English
Dates	06/91-02/19	01/00-04/19
# articles	707	1,774
# unique frames	460	540

Table 1: Datasets overview

Section 4 the final outcome is discussed and the implications of the project are considered.

2 DETECTING ADVERSE MEDIA

2.1 Overview

In the experiment described in this section, we show that statistical supervised learning which utilizes semantic frames as features can be applied to the task of filtering adverse media for due diligence.

We train a logistic regression model that classifies articles as either adverse media or not. Since our data is unlabeled, we use active learning to train the model. The idea behind active learning is that "a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns" (7). This method has been shown to be effective in scenarios where there is plenty of unlabeled data but annotation is expensive and time consuming. The details of our implementation are presented in Section 2.3.

One model is trained on a dataset of news about our use-case company, Mitsubishi Materials, then evaluated on a dataset about another company, Kobe Steel (see Table 1 for an overview of the two datasets). The same procedure is repeated in the other direction as well, i.e. a model which is trained on the Kobe Steel data is evaluated on the Mitsubishi Materials data. The two datasets share one salient adverse media topic: a legal scandal related to data falsification. However, each dataset also contains additional topics that are not shared with the other one (see Table 2). This setup allows us to evaluate how well the models generalize both to similar adverse media topics about a different company, and to topics that have not been encountered during training at all.

The features used to represent the articles are the weighted frequencies (tf-idf) of the FrameNet frames found in them. FrameNet[6] (2) is a lexical database containing more than 1,200 semantic frames; each frame describes a type of event or relation and the participants in it. The intuition behind the choice to use the tf-idf of semantic frames as features is that adverse media might be characterized by certain types of events, e.g. related to legal procedures or deceitful behavior, that are not frequent in the overall data.

[6]https://framenet.icsi.berkeley.edu/fndrupal/
[7]Nexis Uni is a LexisNexis database which offers full text news articles for academic research: https://www.lexisnexis.com/en-us/products/nexis-uni.page

2.2 Pre-processing

We use the NewsReader (8) NLP pipeline to process the text. The pipeline includes *IXA pipes* (1) for tokenization, lemmatization, part-of-speech tagging, syntactic parsing, coreference resolution, word-sense-disambiguation and more. The two modules most relevant for our purposes are the named-entity-detection (NED) and the semantic-role-labeling (SRL) modules.

The NED module detects named entities - such as persons, organizations and locations - classifies them according to their type, and links them to the corresponding DBpedia entries. The module is built on top of DBpedia Spotlight[8], a tool that automatically annotates mentions of DBpedia resources in text.

The SRL module detects PropBank (5) predicates and links them to other lexical resources, including FrameNet, using the PredicateMatrix[9] (3).

The NewsReader NLP pipeline outputs linguistic annotations per document in the *NLP Annotation Format* (NAF) (4). We then parse the NAF files to extract (a) information about the PER and ORG entities mentioned in the document, which is utilized for building graphs (Section 3), and (b) the FrameNet frames of the predicates mentioned in the document, which are utilized as features for machine learning (Section 2.3).

2.3 Active Learning

To train the model, the following procedure is applied:

(1) Select a random sample of articles (N=100) from the dataset and annotate it (adverse media / misc.).
(2) Train a logistic regression model on the labeled data (features: tf-idf of the articles' FrameNet frames).
(3) Run the model on the remaining unlabeled data.
(4) Select a sample (N=20) of articles that the model is least certain about, based on the distance of the data-points to the decision boundary (i.e. the hyperplane separating the two classes).
(5) Annotate the uncertainty-based sample and add it to the pool of labeled data.
(6) Repeat steps (2)-(5) ten times, until N=300 articles are annotated.

This procedure is applied two times: once to the MM dataset and once to the KS dataset. The adverse media topics that were identified during this routine are summarized in Table 2.

The final model available after ten active learning iterations is then evaluated on the other dataset, as described in the next section.

2.4 Results

2.4.1 Evaluation Samples. For each dataset (MM and KS), two evaluation samples are created:

- Active learning sample (N=300). This is the annotated sample that had been obtained during the active learning iterations, as described in Section 2.3.
- Random sample of articles from the dataset (N=300).[10]

[8]https://www.dbpedia-spotlight.org/
[9]http://adimen.si.ehu.es/web/PredicateMatrix
[10]Since the random sample is selected from the whole dataset, including the articles used during active learning, there is a certain overlap between the two sets.

Topic	# articles
MM dataset	
data falsification	65
forced labor during WWII	36
groundwater contamination	2
condos on contaminated soil	2
factory blast	1
KS dataset	
data falsification	115
tax evasion	2
asbestos-related employee death	1
employee embezzlement	1
safety and health violations	1

Table 2: Adverse media topics encountered during annotation

We use these two different evaluation sets because the active learning sample might be unrepresentatively difficult, since it contains the data-points the model was most uncertain about during training. The random sample, on the other hand, is representative of the overall dataset.

We test the performance of the final model (after ten active learning iterations) trained on the MM data on the two samples of the KS data; we test the final model trained on the KS data on the two samples of the MM data.

2.4.2 Quantitative Evaluation. The results of the quantitative evaluation are summarized in Table 3; we focus on the results for class 1 (in bold), i.e. the adverse media items. Overall, the performance of the MM model on the Kobe Steel samples is lower than the performance of the KS model on the Mitsubishi Materials samples. The recall of the MM model is especially low: it retrieves only 55% of the negative articles in the random KS sample, and 31% in the active learning KS sample. The performance of the KS model on the Mitsubishi Materials data is better, with 62-63% recall and 80-82% precision in detecting adverse media both in the random and the active learning MM samples. These differences in performance suggest that the MM and KS datasets might differ in a relevant way; it seems that the KS dataset is more difficult than the MM one. We explore this further in Section 2.5.

2.4.3 Qualitative Evaluation. Although the quantitative metrics are not very high, it is important to note that they are not necessarily the most important ones for the due diligence task. When processing a dataset of news articles about a company, the most important element is to identify all adverse media topics found in it. Once the person performing the due diligence check knows that the company was involved in e.g. data falsification, s/he does not necessarily need to see hundreds of articles on this subject. Therefore, it is more important that the method manages to detect a variety of topics, even if they had not been encountered during training.

In this respect, both models show very good results. We analyzed the true positives detected by the models in the active learning samples to check which adverse media topics (from Table 2 above) the models manage to identify. The MM model manages to detect articles from all five topics encountered in the KS dataset; the KS model manages to detect four out of the five topics encountered in the MM data (all except for the "condos on contaminated soil").

2.5 Analysis of the Results

The quantitative results described in Section 2.4.2 are mixed: while the evaluation on the MM data is promising, the results on the KS data are quite low. This makes it hard to determine whether semantic frames are indeed good features for the task at hand, as we hypothesized. In this section, we explore this issue further.

The numbers (as well as our impression during annotation) suggest that the KS data is noisier in comparison to the MM data. To explore this, we visualize the final annotated samples[11] of both datasets using the t-SNE technique (6). t-SNE allows to visualize high-dimensional data in a two-dimensional space by modeling each high-dimensional object in such a way that similar objects end up close to each other and dissimilar objects end up far away from each other.

The visualization is shown in Figure 2; each article in the annotated sample is a high-dimensional object (the number of dimensions equals to the number of unique frames in the dataset, see Table 1), modeled in two dimensions. In the MM plot, the adverse media articles (orange points) cluster together. More precisely, there are two main clusters in the MM data: the cluster located around x=-5 corresponds to the topic of forced labor during WWII and the cluster located around 0<x<5 corresponds to the data falsification topic. There are only a few adverse media articles that are scattered in other areas of the plot. This shows that in terms of the composition of semantic frames, the adverse media items are indeed similar to each other and distinct from the other articles in the dataset, i.e. correct classification based on semantic frames is possible.

In the KS plot, there is a cluster in the middle of the plot, which corresponds to the data falsification topic; however, there are also many adverse media articles scattered all over, i.e. they are very different from each other in terms of their frames composition. We think that this is a consequence of a certain type of articles prevalent in the KS dataset that have the format of 'daily business news': a list of unrelated items, only one of which is about Kobe Steel. What makes such articles very different from each other is the fact that all the semantic frames found in them are used as features, while the majority of these frames is not related to the adverse media topic and varies from article to article. If this format were excluded from the dataset, we believe that the results would be more similar to what was observed for the MM data; we leave this investigation for future work.

To conclude, the plots support our hypothesis that semantic frames are sensible features for adverse media detection. It seems that adverse media articles, regardless of their topic, tend to be similar to each other in terms of their semantic frames. This ability to generalize beyond a specific topic is further supported by the results reported in Section 2.4.3: the model trained on the MM data correctly retrieves adverse media not only about data falsification, but also about a range of topics it did not encounter in training,

[11] The final annotated sample is the union of the active learning and the random samples (see Section 2.4.1). For MM, the size of the final sample is N=473; for KS, N=557.

Model	Test set	Class	Precision	Recall	F1-score	Support
MM_10	KS: active learning sample (N=300)	0	0.65	0.89	0.75	177
		1	**0.67**	**0.31**	**0.42**	**123**
MM_10	KS: random sample (N=300)	0	0.90	0.95	0.92	242
		1	**0.73**	**0.55**	**0.63**	**58**
KS_10	MM: active learning sample (N=300)	0	0.82	0.91	0.86	194
		1	**0.80**	**0.63**	**0.71**	**106**
KS_10	MM: random sample (N=300)	0	0.91	0.97	0.94	240
		1	**0.82**	**0.62**	**0.70**	**60**

Table 3: Quantitative Evaluation Results (class 1: adverse media)

like tax evasion, embezzlement, safety and health violations, and an asbestos-related death. Similarly, the model trained on the KS data correctly retrieves adverse media not only about data falsification, but also about forced labor during WWII, groundwater contamination, and a factory blast. These results suggest that if a model is trained on data containing a few representative topics of interest, it will be able to generalize well to many different adverse media topics.

Moreover, using semantic frames as features might make the models more transparent and interpretable in comparison to token-based models. Table 4 shows which ten features (frames) have the highest coefficients in the two models. In the KS_10 model, we find the frames `Reveal_secret` (example predicates: *leak*, *admit*), `Forging` (e.g. *falsify*), `Inspecting`, `Criminal_investigation`, `Try_defendant` and `Verdict` (e.g. *convict*). These frames are clearly related to the type of content our method aims to detect. The relevance of the top-ten frames of the MM_10 model is less transparent; however, it is important to note that the full picture depends on the combination of all features and their relative weights, so more in-depth analysis (which is beyond the scope of this paper) is required.

MM_10 model	KS_10 model
Being_in_effect	Reveal_secret
Intentionally_create	Forging
Manipulate_into_doing	Inspecting
Locating	Abounding with
Choosing	Criminal_investigation
Becoming_aware	Try_defendant
Quitting	Manufacturing
Forging	Verdict
Work	Attaching
Participation	Research

Table 4: Ten features with the highest coefficients

To sum up, we show evidence that statistical supervised learning which utilizes semantic frames as features can be applied to the task of filtering adverse media for due diligence. Further experimentation with different datasets, different adverse media topics and different machine learning algorithms is needed to draw more

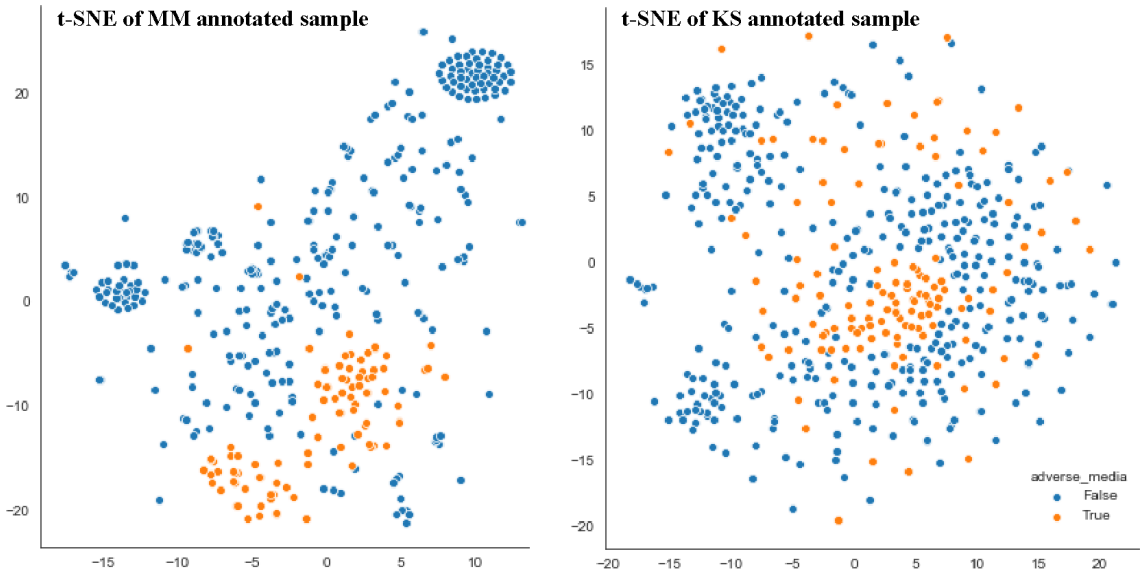

Figure 2: t-SNE plots of the final annotated samples of the MM (left) and the KS (right) datasets

general conclusions and to build a production-ready application. Our work provides the first proof-of-concept step in this direction.

3 GENERATING KNOWLEDGE GRAPHS

A knowledge graph represents a knowledge base that allows to query over information gathered from a variety of sources. In the present use-case the knowledge graph would aid to visually represent the direct links between Mitsubishi Materials and other people or companies identified in the adverse media articles and the structured data sources (DBpedia and Offshore Leaks Database), and more importantly by linking all these sources together the knowledge graph would allow to identify indirect links which would expand the knowledge base of either source alone. The knowledge graph was built in a graph store, GrapDB 8.8.1.

3.1 Newspaper Entities

In total 2458 entities are extracted from the adverse media articles from Mitsubishi Materials data set. Out of these 2458 entities 247 are unique. The large amount of repetitions is due to the fact that usually multiple newspapers report the same event, hence, there is an overlap in the data. It was chosen to only use entities that have a DBpedia reference, as then the identification of the entities would be unambiguous, more information could be extracted via DBpedia, and it would assure that the entities can be linked to other data sources. Consequently, 149 unique DBpedia links were extracted, out of which 123 were imported in the GraphDB database. The decrease in the final import was because some of the entities identified were irrelevant, such as newspaper company names, therefore they were filtered out. Lastly, all entities mentioned in the same news article were explicitly linked together in the graph via a reciprocal link, news:mentioned (Figure 3).

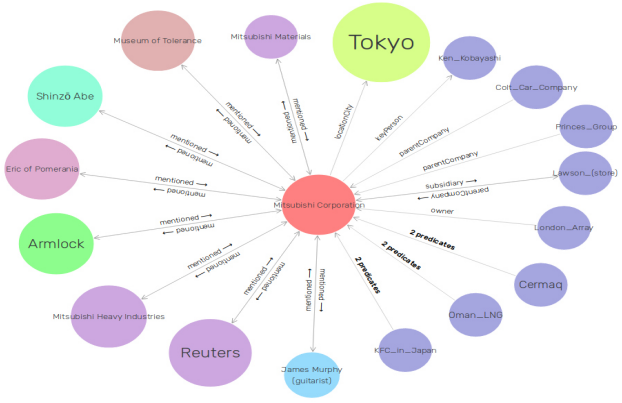

Figure 3: News Graph

Figure 3 represents a small part of the information stored in the graph for the entity 'Mitsubishi Corporation'. The entities on the right (purple) represent the information that was obtained from DBpedia, while the entities on the left represent some of the entities that were identified in the same article as Mitsubishi Corporation (see the link 'mentioned'). As the information for all the entities extracted from the news search was expanded via DBpedia, all the entities on the left can be further expanded, thus providing new information about connections to other people and companies, that was not known before from the news search (Figure 4).

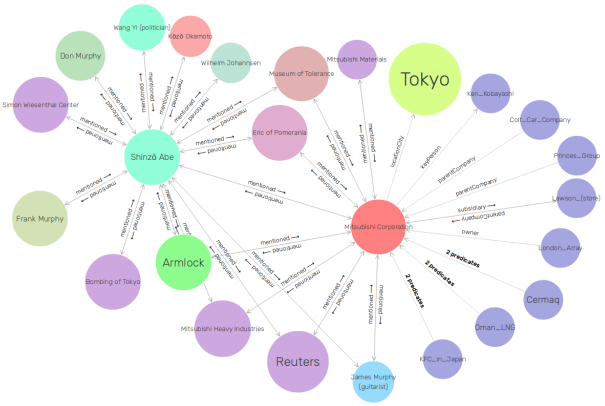

Figure 4: Expansion of one entity

3.2 Offshore Leaks Database

The data set contains more than 785 000 offshore entities that have been part of the Paradise Papers, the Panama Papers, the Offshore Leaks and the Bahamas Leaks investigations. The value of this database is that it exposes companies and people involved in offshore investments (international tax fraud). Hence, the offshore entities can be a person or a company, and the data set specifies the connections between them. In total there are 3 entity types described: Officers, Intermediaries and Entities. The relationships between these entities are summarized in Figure 5

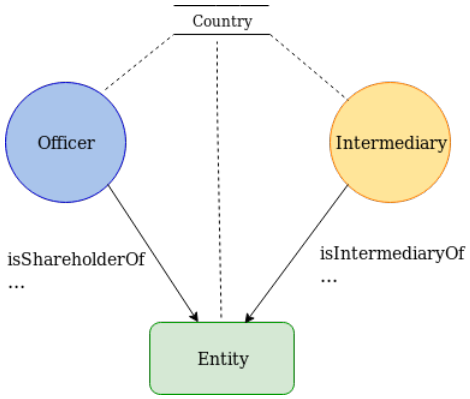

Figure 5: Relationships between Offshore Leaks entities

3.2.1 Data Investigation. The database was imported in GraphDB according to Ontotext manual established in 2016 [12]. Even though the Offshore Leaks Database contains a lot of information, thus far its use has been limited to queries only over the data itself. This is due to the fact that apart from the country location of an

[12] https://github.com/Ontotext-AD/leaks

	Linked Entities	Officers	DBpedia	Intermediaries	DBpedia
Mitsubishi Corporation	Asia Group Investments Limited	42	0	1	0
	Energi Mega Pratama Inc.	16	2	2	1
	CP Secure International Holding Limited	66	4	1	1

Table 5: Linked entities to the use-case, and the investigated relations of these entities

entity there is no external endpoint to which the entities would be linked to, hence, in the current representation the data cannot be linked to other databases. To overcome this limitation it was chosen to try to identify the entities on DBpedia and if found, create a URI based on the DBpedia page. As this identification process was executed manually, the location of the entities was limited to a certain country to limit the amount of entities. The chosen country was Japan as it is the registered location of the use-case 'Mitsubishi Corporation'.

The outcome of this search resulted in 28 Entities, 47 Intermediaries, and 899 Officers, whose country location was specified as Japan. As the entity type 'Entities' represent an endpoint of a relation link (Figure 5), it was chosen to start the identification on DBpedia with these entities. Out of the 28 Entities identified none were found on DBpedia, hence the Officers and Intermediaries of these Entities were examined instead. The results of this query amounted to 36 Officers and 28 Intermediaries, out of which 4 and 1 were identified on DBpedia. However, this search lead to a dead end, therefore, the Officers identified in the first step were examined further. This was done by running multiple SPARQL queries with specifying the subject as type Officer, object as type Entity, and varying the predicate per query. The outcome of this approach brought the following results: 'Mitsubishi Corporation' was identified as an officer in the Offshore Leaks Database.

As the main goal was to identify relations that 'Mitsubishi Corporation' has with other companies/people, the query was updated by setting a fixed subject, the Officer id of 'Mitsubishi Corporation', and extracting all related entities to it. This search identified 3 entities: 'Asia Group Investments Limited', 'Energi Mega Pratama Inc.', and 'CP Secure International Holding Limited' (Figure 6 (the 3 entities at the bottom)). Nonetheless, none of these entities could be identified on DBpedia, hence the Officers and Intermediaries of these 3 entities were examined as well. In table 5 the outcome is summarized.

The further investigation did not reveal any overlap between the three entity Officers or Intermediaries (Figure 7). Hence, as no further information could be extracted from the graph it was chosen to search the 3 identified entities on the news, as this might provide additional information, that is not in DBpedia or the Offshore Leaks Database. The adverse news article search of 'Asia Group Investments Limited', 'Energi Mega Pratama Inc.', and 'CP Secure International Holding Limited' did not reveal any additional relevant information. Therefore, a 'dead end' was reached and with the present time limitations it was chosen to stop the further expansion of the graph, and evaluate the present results.

4 CONCLUSION

We introduce a novel approach that assists the human expert with performing a due diligence investigation. Our system (Figure 1)

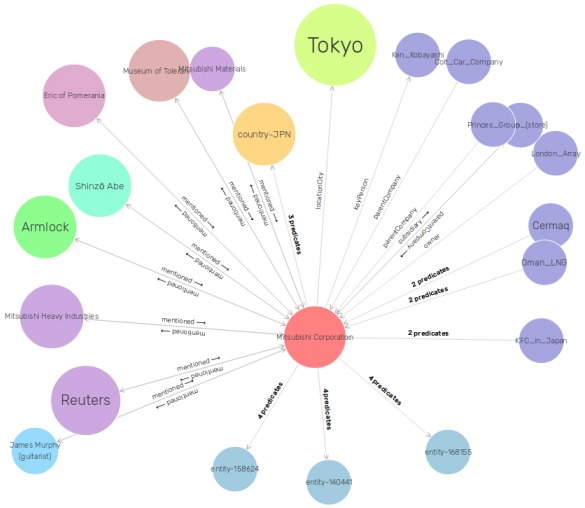

Figure 6: Identified entities in the Offshore Leaks Database

filters out adverse media articles from a large collection of news items, interlinks the information extracted from these articles with multiple structured data sources, and outputs a graph representation of the knowledge obtained from both the structured and the unstructured data. This setup has the potential to generate new insights by discovering unknown indirect connections between the entities in the graph. Since our pipeline can be applied iteratively (by inputting newly discovered entities to the news search), the network of the original seed company can be expanded to more and more remote connections, depending on the end user's needs.

In this paper, we present the first proof-of-concept experiment to show the promise of this approach. We provide evidence that statistical supervised learning which utilizes semantic frames as features can be successfully applied to the task of adverse media detection. We also show that using the active learning method allows us to train a reasonably-performing model with a very limited annotation effort (300 documents). Based on the insights obtained in our experiment, we believe that a model trained on data that contains a few representative adverse media topics will be able to generalize well to other topics, which have not been encountered in training.

Moreover, we explore the potential of interlinking various data sources in a graph representation. For the use-case used in this paper we could not demonstrate the full strength of this method since many of the extracted entities could not be identified in DBpedia or in the news database. A possible solution is to incorporate additional structured data resources into the system, e.g. legal databases and company registries.

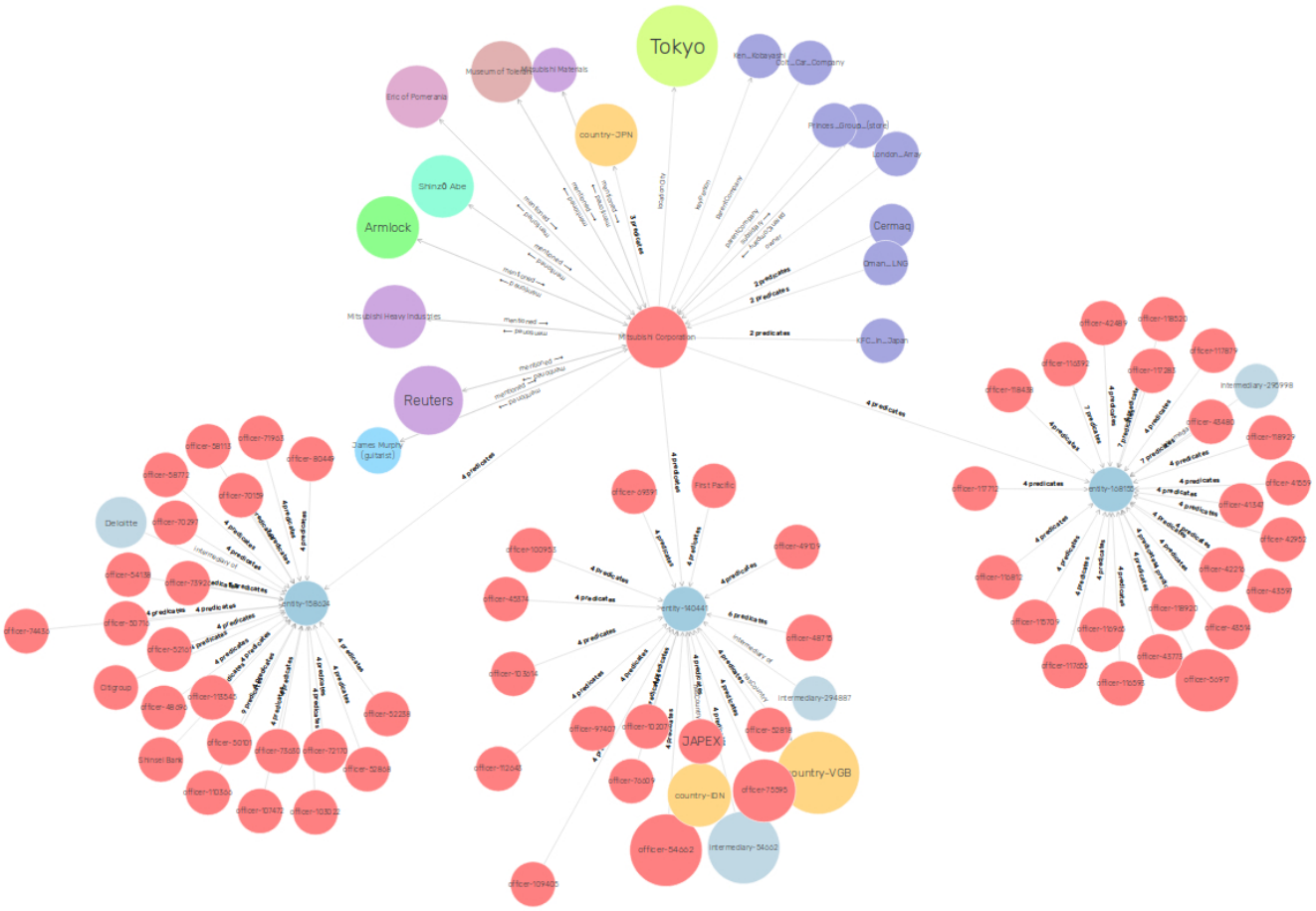

Figure 7: Extracted Officers and Intermediaries of the 3 identified entities

As this was the first attempt to implement the method, future research should focus on evaluating the approach. For a full evaluation, we would need to use a known case of a suspicious indirect connection between two entities, which is not explicitly mentioned in the media and in the structured data. Finding this connection (i.e. generating new information which does not exist in any of the separate sources) would demonstrate the full promise of our approach. Despite these limitations and even in its current state, we believe that our pipeline can assist and enhance the manual due diligence procedure.

REFERENCES

[1] Rodrigo Agerri, Xabier Artola, Zuhaitz Beloki, German Rigau, and Aitor Soroa. 2015. Big data for Natural Language Processing: A streaming approach. *Knowledge-Based Systems* 79 (2015), 36–42.

[2] Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 86–90.

[3] Maddalen Lopez De Lacalle, Egoitz Laparra, and German Rigau. 2014. Predicate Matrix: extending SemLink through WordNet mappings.. In *LREC*. 903–909.

[4] Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, German Rigau, Willem Robert van Hage, and Piek Vossen. 2014. NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*. 9–16.

[5] Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank.. In *LREC*. Citeseer, 1989–1993.

[6] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[7] Burr Settles. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

[8] Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, et al. 2016. NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems* 110 (2016), 60–85.