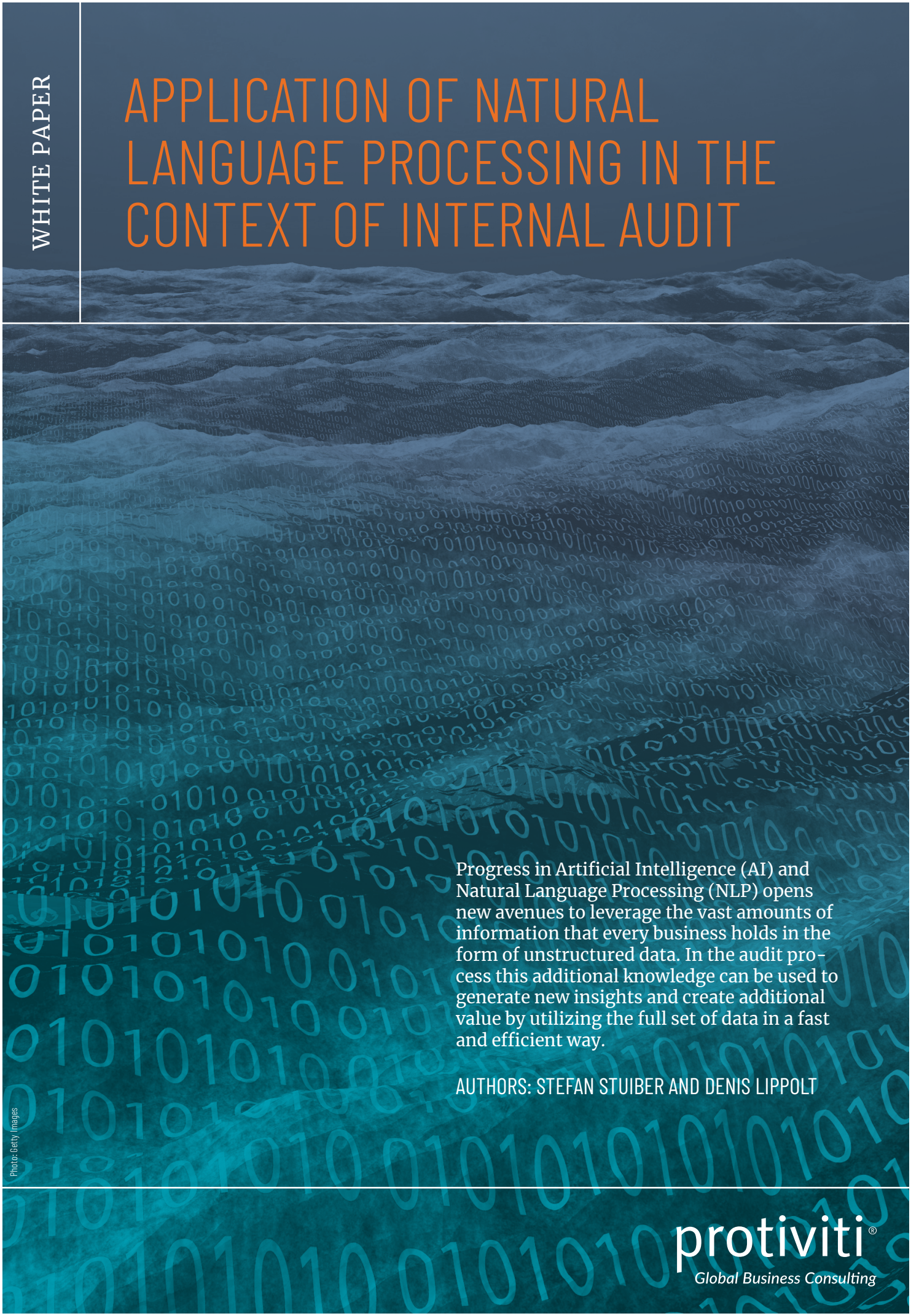
WHITE PAPER

APPLICATION OF NATURAL LANGUAGE PROCESSING IN THE CONTEXT OF INTERNAL AUDIT



Progress in Artificial Intelligence (AI) and Natural Language Processing (NLP) opens new avenues to leverage the vast amounts of information that every business holds in the form of unstructured data. In the audit process this additional knowledge can be used to generate new insights and create additional value by utilizing the full set of data in a fast and efficient way.

AUTHORS: STEFAN STUIBER AND DENIS LIPPOLT

Photo: Getty Images

protiviti®
Global Business Consulting

INTRODUCTION

Natural Language Processing (NLP) is a subdiscipline of artificial intelligence technologies at the intersection of linguistics and computer science. It encompasses a range of computational methods for analyzing and representing naturally occurring texts in a linguistic analysis to recreate human-like processing of various tasks and applications [1]. NLP can therefore be described as a machine's ability to understand the human language by processing interactions between humans and computers in natural language. NLP is the centerpiece of many common applications: automated translations; search queries that go beyond simple keywords, where information is provided as a response to a question; or voice controlled personal assistants like Siri and Alexa.

NLP AND AUDIT

NLP technologies can be used to access and leverage the vast amount of information that every company deals with in the form of unstructured data. Such data accounts for approximately 90% of all data in a company [2] and contains texts, images, videos, and audio files. This alone illustrates the potential and the need for automated analysis. This information can provide useful insights, especially in relation to internal audit processes.

Internal audit forms the "third line" of defense in the "Three Lines (of Defense)" model. The model provides a framework of an effective governance system for organizations, with clearly defined roles and responsibilities for each of the three lines. The "first line" is assigned to the operational management; it is responsible for the evaluation, control and reduction of the risks, and is accountable for the implementation of controlling mechanisms along the whole value-added chain. Process owners as part of the second line hold the responsibility for the processes, so that potential risks can be prevented, recognized, and corrected at an early stage. In this way the second line oversees, controls, and supports the management in the risk control function. It is deployed by management and releases rules and guidelines to implement the risk control strategy. The "second line" works closely with the management and reports to it. In order to assure that the risk control and operation management functions perform effectively and appropriately, an independent and objective "third line" is needed, and thereby formed by internal audit. The risk management is evaluated and assessed. Findings are reported to the top management, and transparently communicated to both the first and second lines. During the audit, business processes must be analyzed thoroughly, making it necessary to gather the complete information from all relevant processes.

Manual content analysis for the audit has the advantage of a high-quality analysis with high precision and granularity. But there are two severe structural problems inherent in manual analysis. Firstly, it is very labor intensive and expensive, which usually prevents an analysis of the full available data set. As a result, a small sample size is selected with limited generalizability and statistical power, where important information or high-risk elements can be overlooked. Secondly, a potential bias could be introduced by a sample selection process performed by auditors, which can limit replicability. With an automated approach to content analysis, the full population of data can be accessed quickly and efficiently – the risk of a biased sample selection process is eliminated. Moreover, by utilizing the full population, many more insights can become available that would

» By utilizing the full population, many more insights can become available that would not be accessible from a sampled population or the structured data alone. «

PETER GRASEGGER, MANAGING DIRECTOR

» With NLP the full population of data can be analyzed to gain additional insights. «

DENIS LIPPOLT, DIRECTOR

not be accessible from a sampled population or the structured data alone. Additional insights can be created by combining NLP with an unsupervised learning approach, e.g., an isolation forest [3], which could provide additional information on high-risk elements within the population.

For an audit, text documents such as reports, bills, receipts, contracts, policies, emails, and meeting transcripts are crucial. NLP analysis has been successfully applied to a range of these documents. The authors of [4] reported the application of NLP to extract detailed information from physical leasing contracts to fulfill changed IFRS 16 requirements for the reporting of operating leases. In a case study for an international technology group, the manual effort was reduced by a factor of five using NLP, and guidelines for the setup of an automated extraction process were developed.

NLP was used in [5] to construct risk profiles from annual 10-K filings; the similarity of sentences was calculated with a predefined list of risk words and then each sentence was categorized into four risk classes, according to the highest similarity score. Afterwards the risk classes for all sentences were combined to create a risk profile for the complete filing, which the authors then validated using audit options and internal controls. The authors furthermore compared the number of sentences relating to risks for companies with and without Internal Control Material Weaknesses (ICMW) and found that companies with ICMW present more risk-related sentences in their filings. Using NLP, it was possible to show that problems in the internal control mechanisms of companies are reflected in the language and style of their annual filings.

An analysis of approximately 300,000 emails enabled the authors of [6] to demonstrate that linguistic models can identify deceptive emails. These emails showed a reduced frequency in the use of first-person pronouns and exclusive words; and showed a higher frequency of words associated with negative emotions, as well as words which call for action.

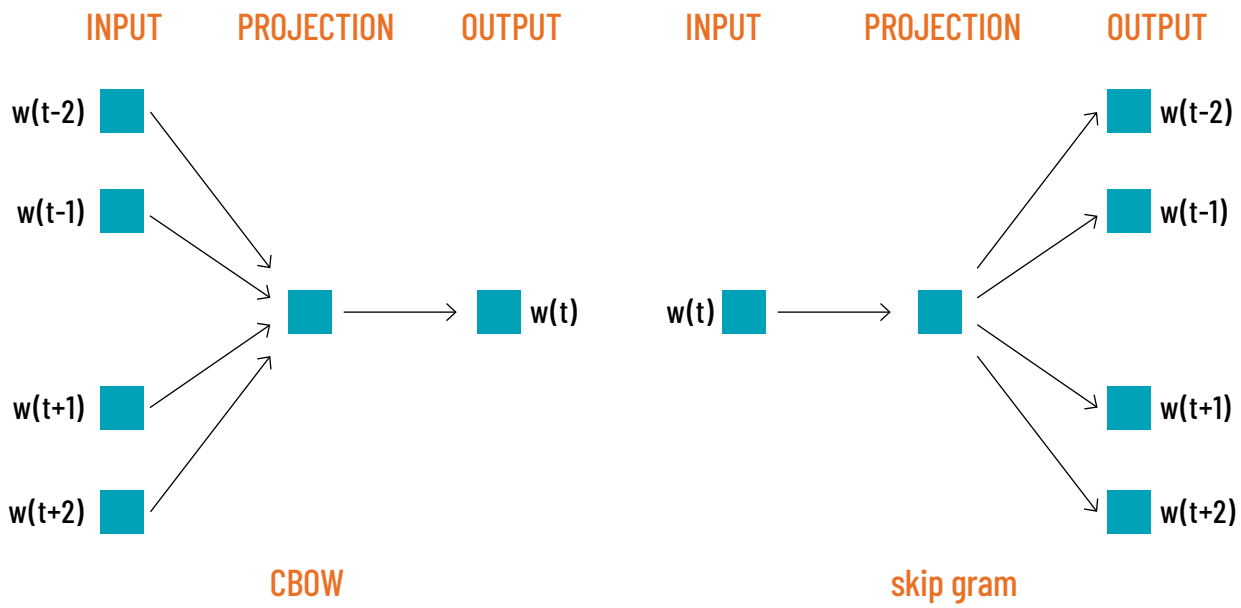
Earning conference call transcripts have been investigated using sentiment analysis by [7]. The sentiment score of the text, as well as the confidence score of the emotion joy, are added as additional predictors to the model to improve the prediction accuracy for ICMW.

In [8], the authors found that the qualitative narrative content of annual reports contains information that is useful for detecting fraud, which is not captured by the financial numbers in the companies' structured data. Systematic differences in communication and writing style indicate fraudulent reports. Their linguistic model predicted fraudulent behavior with an accuracy of almost 90%.

Similarly, these technologies can be applied to further problems that arise in the context of internal audit. In the audit of credit decision and credit limit review documentation a large amount of information is available only as unstructured data. This limits the ability of a sample-based approach in audit to capture all aspects of these processes. With NLP the full population of data can be analyzed, and information can be extracted from the texts within these documents to gain additional insights. Guidelines can be extracted from policy documents, and individual customer files can then be compared to the rules automatically. The same idea can also be applied to insurance claims filed by customers, to have their damages covered. Know-your-customer processes can be reviewed by analyzing the unstructured data with NLP tools, e. g., customers can automatically be categorized into risk classes or as politically exposed persons, according to the information provided in application forms.

These examples highlight the additional value that lies in unstructured data and the analysis of linguistic information within the audit context. In the following section, we present the general steps necessary to preprocess unstructured textual data, to prepare them for a variety of automated NLP tasks, such as information extraction, classification, or clustering. Our example is a set of 1,300 publicly available CVs, which could be successfully processed.

FIGURE 1: Illustrations of the CBOW (left) and continuous Skip-gram (right) networks. CBOW predicts one word from its context, whereas skip-gram predicts the surrounding words from the current single word.

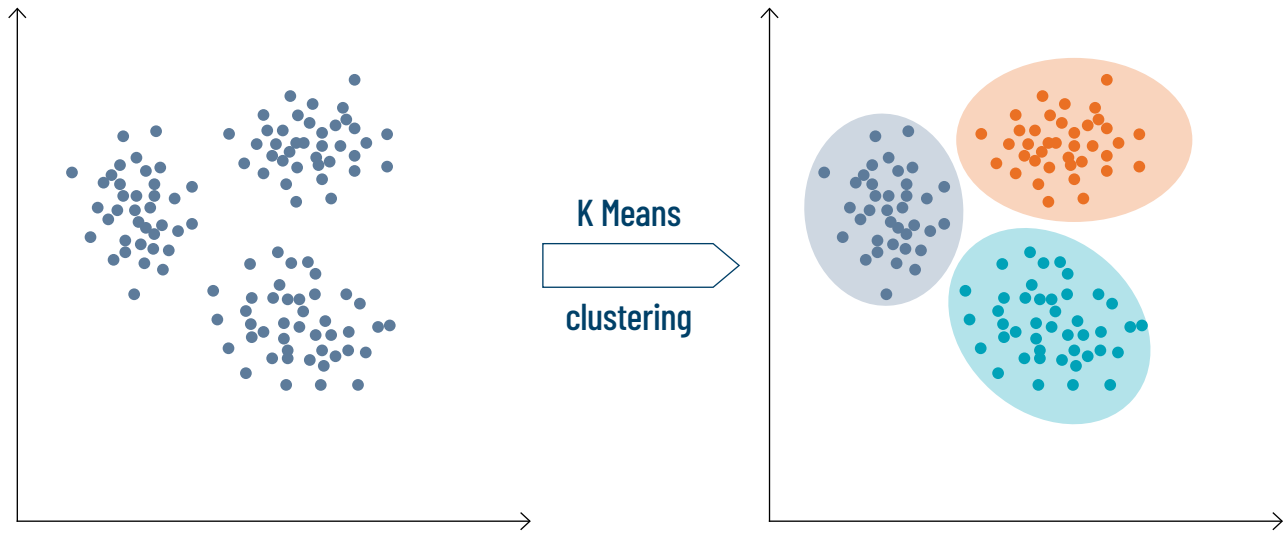


NLP METHODOLOGY

The first step in natural language processing is to extract, transform and load the data, converting it into a form that can be processed further by a machine. For unstructured text data from e.g., pdf files or images, optical character recognition (OCR) is a widely used deep learning application to extract the textual content. However, the result of the OCR process might still contain typos, non-text symbols and recognition mistakes, and requires further processing. Conventional approaches include looking up words in dictionaries or employing statistical language models. Files containing different formats of text, like invoices, where numbers are listed in a table, pose another challenge for OCR algorithms, and require additional computer vision or machine learning tools to be correctly transformed into a machine-readable format. Recordings of conference calls can be transcribed manually, or a speech-recognition software can be used to accomplish the task. Speech recognition algorithms were traditionally based on statistical tools like Hidden Markov Models (HMMs), but recently, deep neural networks have emerged as the leading technology [9].

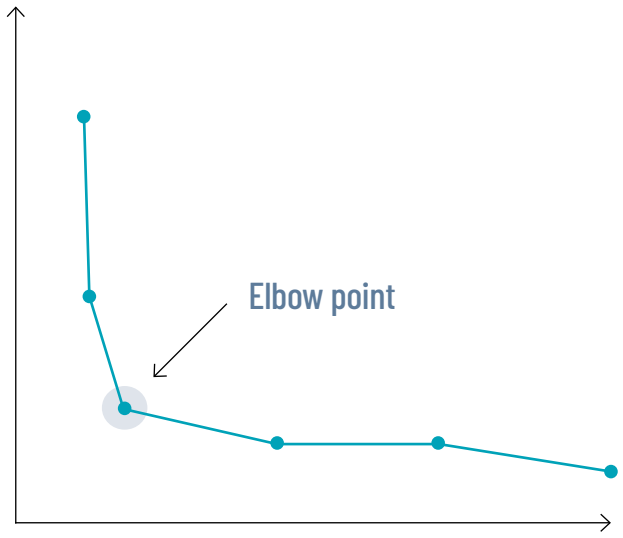
With text in digital form, the preprocessing for NLP can be started. Words like articles, prepositions and conjunctions are called "stop words" and can be excluded from the analysis, since they occur very often and contain limited information. For the remaining words, the next step is stemming or lemmatization. These are two mechanisms to reduce a word to its stem – the part of the word responsible for its meaning. Stemming is rule-based: an algorithm follows a set of steps to remove suffixes from words. One implementation is for example the Porter Stemmer [10]. Lemmatization uses a different approach by first applying a tagger to each word in the sentence to identify the intended part-of-speech for the word, and then using a dictionary to reduce the word to its lemma. The following example illustrates the difference. A stemmer might reduce the word "saw" to only "s," by cutting off the ending "aw," assuming it was the inflectional ending. Lemmatization would tag the word "saw" either as a verb or as a noun, depending on the sentence, and then return the lemma "see" or "saw" accordingly. Both algorithms can greatly reduce the size of the index, but this reduction has the potential to also remove information from the text or obscure meaning. The remaining words form the vocabulary for the text.

FIGURE 2: K-Means clustering finds structures in the unstructured document representations, and groups them together into a predefined number of clusters.



After preprocessing, the text is ready to be embedded. Embedding refers to representing words (or text) in a way computers can handle and analyze, typically in the form of real-value vectors. One possible, very basic embedding technique is a one-hot encoding. Here, a vector is created with a length the size of the vocabulary. A word is then numerically represented by a vector with a value of one at a specific index, and zeros at all other indices. Another basic embedding is also the bag-of-words model; the number of occurrences of each word in the whole text is counted, and a sentence is represented as a vector, containing the number of occurrences for each word in that sentence. More advanced embeddings also encode the meaning of a word, in the sense that words with similar meaning are represented by vectors close to each other in the vector space. One network to generate these kinds of mappings is the Word2Vec created by [11]. It uses the cosine similarity from the scalar product between two vectors as the distance function. In this way, the embedding can capture semantic patterns. Algebraic operations on the vectors are able to calculate relationships like "King"-"Man"+ "Woman" = "Queen." For Word2Vec, two different model architectures are available to produce distributed representation of the words: Continuous Bag of Words (CBOW) and continuous Skip-gram (SG) **(see also Figure 1)**. Distributed representation means that for each word vector some dependence on the surrounding words is encoded. For CBOW the order of the words does not contribute to the embedding – only a window of surrounding words provides the context to predict a word. In the SG architecture it is the other way around; the current word is used to predict the surrounding window of context words. The Word2Vec model has been extended to Doc2Vec, where an additional input parameter for the paragraph or document is added to the input vector. This parameter can serve as a memory to capture semantics of more context than just surrounding words. Documents could, for example, be tagged with the author's name. This information is included in the word vectors and can be used for classification. Word2Vec and Doc2Vec allow for high-quality vector representations of words in simple network architectures, with low computational costs. Therefore, much larger data sets are accessible. To make full use of these vector representations, the networks should be trained on very large data sets. However, a major disadvantage of these models is that they do not support out-of-vocabulary words – they cannot handle words they have not been trained on. Furthermore, they create static representation of words, and cannot distinguish homonyms that appear in different contexts within the same corpus. One approach to overcome these disadvantages are Bidirectional Encoder Representations from Transformers (BERT) language models [12]. They provide

FIGURE 3: Elbow method to heuristically find the optimal number of clusters for K-Means clustering. The variance is plotted against the number of clusters. A sharp bend (Elbow point) indicates that more clusters will not reduce the variance much further.



embeddings not for single words, but for entire sentences. Pretrained BERT models can be adapted to specific corpora with only one additional output layer, which makes them applicable to a wide range of tasks (e.g., FinBERT which has been trained on 4.9 billion tokens from financial documents in approximately two days on four GPUs [13]).

Word vectors like this also enable mathematical text analysis. For example, texts can be clustered with the unsupervised K-Means algorithms [14] to reveal additional information not accessible from only single documents. In K-Means clustering **(see Figure 2)**, a data set is segmented into a predefined number of groups. To determine a useful number of clusters, the heuristic elbow method **(see Figure 3)** can be used. A measure of the variance is calculated and plotted as a function of the number of clusters. At some point, adding more clusters will not decrease the variance significantly, and an elbow forms in the graph. This value for the number of clusters provides a good compromise between over- and underfitting the data set. In the algorithm, the groups are formed automatically, in a way that the distance to the center of each cluster is minimized. Since the chosen word vectors contain meaning and similarity of words, these vectors are close to each other and will also be clustered together by the algorithm. The quality of the clustering process can be estimated by the inertia value: the sum of all squared inner-cluster distances or the silhouette values, which compare the distance of a data point within its cluster to the data point's distance to the points in the other clusters.

The word representation vectors usually have a very high dimension. A tool for dimension reduction and features separation is Principal Component Analysis (PCA) [15]. High dimensional data is projected into a vector space with a lower dimension, losing as little information as possible and combining possible redundancy in the data. The resulting principal components are ordered, with the first one containing the largest amount of variation. Another similar technique is factor analysis, which allows more intuitive understanding of the components. There are some subtle differences to PCA which are beyond the scope of this work. For visualization, the data dimension can be further reduced with t-distributed stochastic neighbor embedding (t-SNE) [16]. Higher dimensional data is mapped to two or three dimensions by learning a probability distribution for the similarity of points on the map, according to their probability distribution in the higher-dimensional space.

All of the 1,300 CVs, with the exception of two (due to faulty pdfs) could be processed completely automatically within minutes. The text was extracted, converted to a numerical representation, and the documents were clustered without manual intervention. With this prototype, the whole NLP chain could be applied to have the full information available, which would provide insights, and could be used for business decisions.

CONCLUSION

NLP offers the possibility to process thousands of documents of unstructured data in a short time and extract valuable information. By using this information, the audit process can be accelerated and made more thorough, since the full data can be included and additional knowledge can be gained. These advantages would not be accessible from a smaller sample. Future audits will profit greatly from the advancement in NLP and its application in the audit context. For the results reported in the literature NLP benefits have been achieved by external teams on documents that were publicly available. Application of NLP methods during internal audit exercises, allows addressing problematic points and remediating them quite earlier in the process. The audit process itself can become more comprehensive, faster, and more efficient with the help of NLP algorithms.

REFERENCES

[1] *E. D. Liddy, Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. 2001.*

[2] *G. Schumann and J. M. Gómez, "Natural Language Processing in Internal Auditing – a Structured Literature Review". AMCIS 2021 Proceedings.*

[3] *Isolation Forest is a widespread high-performance unsupervised learning algorithm, that enables detection of anomalies. More details can be found in the paper "Internal Audit Application of Machine Learning Sample Selection", https://www.protiviti.com/sites/default/files/2023-04/white-paper_internal_audit_applications_of_machine_learning.pdf*

[4] *J. H. Mayer et al., "Towards Natural Language Processing: An Accounting Case Study", 41st International Conference on Information Systems Proceedings (2021), https://www.rcw.wi.tu-darmstadt.de/media/bwl4/forschung_9/kompeten-zzentrum_1/20200927_ICIS_1383_NLP_track_23_final_version_letter_size.pdf .*

[5] *M. Lui et al., "Textual Analysis for Risk Profiles from 10-K filings", The CPA Journal (2020), https://www.cpajournal.com/2020/07/15/textual-analysis-for-risk-profiles-from-10-k-filings/ .*

[6] *P. S. Keila et al., "Detecting unusual and deceptive communication in email", Proceedings of the 2005 Conference of the Centre for Advanced Studies on Collaborative Research (2005). IBM Press, pp. 17-20.*

[7] *T. Sun, "Deep learning applications in audit decision making". PhD Thesis, The State University of New Jersey, 2018.*

[8] *S. Goel et al., "Can linguistic predictors detect fraudulent financial filings?" Journal of Emerging Technologies in Accounting 7 (1) (2010), pp. 25-46.*

[9] *U. Kamath et al., "Deep Learning for NLP and Speech Recognition", Springer International Publishing 2019.*

[10] *M. F. Porter, "An algorithm for suffix stripping", Program, 14 (3) (1980), pp. 130-137.*

[11] *T. Mikolov et al., "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781 (2013).*

[12] *J. Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint, arXiv:1810.04805 (2018).*

[13] *Yang, Yi, et al., "Finbert: A pretrained language model for financial communications", arXiv preprint, arXiv:2006.08097 (2020).*

[14] *J. MacQueen, "Some methods for classification and analysis of multivariate observations", Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (1967), pp. 281-297.*

[15] *M. E. Tipping and C. M. Bishop. "Mixtures of probabilistic principal component analyzers", Neural computation 11.2 (1999), pp. 443-482.*

[16] *L. van der Maaten und G. Hinton, "Viualizing data using t-SNE", Journal of Machine Learning Research, Vol. 9 (2008), pp. 2579-2605.*

CONTACT US!

DENIS LIPPOLT
Director
+49 172 698 30 48
denis.lippolt@protiviti.de

PETER GRASEGGER
Managing Director
+49 173 653 8922
peter.grasegger@protiviti.de

www.protiviti.de