Variations in Word Usage for the Financial Domain

Syrielle Montariol[1,2,4*] , **Alexandre Allauzen**[3] , **Asanobu Kitamoto**[4]
[1] Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France.
[2] Société Générale, Paris, France
[3] ESPCI, Université Paris Dauphine - PSL, Paris, France
[4] National Institute of Informatics, Tokyo, Japan
syrielle.montariol@limsi.fr, alexandre.allauzen@espci.fr,
kitamoto@nii.ac.jp

Abstract

Natural languages are dynamic systems; the way words are used vary depending on many factors, mirroring the divergences of various aspects of the society. Recent approaches to detect these variations through time rely on static word embedding. However the recent and fast emergence of contextualised models challenges the field and beyond. In this work, we propose to leverage the capacity of these new models to analyse financial texts along two axes of variation: the diachrony (temporal evolution), and synchrony (variation across sources and authors). Indeed, financial texts are characterised by many domain-specific terms and entities whose usage is subject to high variations, reflecting the disparity and evolution of the opinion and situation of financial actors. Starting from a corpus of annual company reports and central bank statements spanning 20 years, we explore in this paper the ability of the language model BERT to identify variations in word usage in the financial domain, and propose a method to interpret these variations.

1 Introduction

It is well know that all languages gradually evolve over decades and centuries, mirroring the evolution of the society. However, variation in word usage is not limited to long-term evolution. Many fine-grained variations can be found at a smaller scale. One the one hand, in the short term, the usage of a word can vary in response to sudden events that do not necessarily alter its meaning, but can momentarily change the way it is used. On the other hand, the usage of a word can vary depending on the person that uses it: several dimensions (geographical, cultural) can lead communities to use words in a different way depending on the local interests and concerns. These two kinds of variations are called *diachronic* (through time) and *synchronic* (across any other dimension than time).

In the financial domain, detecting the variations in word usage through time can lead to better understanding of the stakes and concerns of each time period [Purver *et al.*, 2018]. In a synchronic way, many dimensions can be observed: how the words are used depending on the business line, the country of origin, the company or organisation that produces the document... This way, the opinions, behaviour and preoccupations of the writer can transpire through its specific usage of words. This information can be useful to financial analysts to better understand the variations of concerns and viewpoints of financial actors (for example, by analysing text from the regulatory authorities), identify the impact of an event on different actors through time (using high temporal granularity data sources), or analyse the evolution of a crisis.

In other words, we look for *weak signals* through the scope of word usage change. A weak signal is an element observed from data that has ambiguous interpretation and implication, but may be of importance in the understanding and prediction of events (present or future). In the financial domain, any change in strategy, emerging concern or unusual event linked to a financial actor can be a weak signal; identifying relevant weak signals and interpreting them is an extremely challenging task.

In this paper, we study word usage change as a potential signal of evolution in the situation and opinion of a financial actor. When an analyst reads a set of financial documents, the diachronic and synchronic variations in word usage are not immediately visible. But they might reveal valuable information, if they can be detected and interpreted. For example, it can be shown that the connotation of the vocabulary used by central banks in their reports and statements is strongly influenced by the economic situation [Buechel *et al.*, 2019], despite the sensitivity of their position.

As a growing amount of historical textual data is digitised and made publicly available, automated methods of Natural Language Processing (NLP) emerge to tackle the diachronic aspect of this task. The models usually rely on static word embeddings such as Word2Vec [Mikolov *et al.*, 2013] which summarise all senses and uses of a word into one vector at one point in time. This prevents the model from detecting more fine-grained variations of word usage (polysemy) according to its various contexts of occurrences. To tackle this problem, a new set of methods called contextualised embeddings has appeared recently. They allow to represent words at the token level by relying on their context. Several pre-trained language models (BERT [Devlin *et al.*, 2019], ELMO [Peters *et al.*, 2018]...) have appeared for this purpose in the past

*Contact Author

two years. We rely on the model BERT, as Wiedemann *et al.*[2019] show its superiority in disambiguating word senses compared to other contextualised embeddings models.

In this paper, we use BERT to determine in a fine-grained way the different kinds of use of a word and the distribution of these uses in a financial corpus. Our goal is to analyse financial texts in a diachronic and synchronic way, as a preliminary investigation to address the following questions:

In a synchronic way, what do word usages reveal about the opinion and behaviour of different financial actors? In a diachronic way, what does it says about their evolution? Can it allow to better understand past and ongoing events through the scope of word usage change?

The key points of this paper are:

1) Studying word use variations across any dimension in the financial domain (e.g. time, business line, financial actor).

2) Proposing a method to measure and interpret the variations of a word usage across a dimension.

The model and the pipeline are described in section 3. The experiments in section 4 are made on a corpus of annual company reports and a corpus of central bank statements, both spanning two decades, described in section 4.1.

2 Related Work

Before the generalisation of word embeddings, measuring diachronic semantic change used to rely on detecting changes in word co-occurrences, and on approaches based on distributional similarity [Gulordava and Baroni, 2011].

A more recent set of methods rely on word embeddings [Mikolov *et al.*, 2013] and their temporal equivalent, diachronic word embeddings models. They rely on the assumption that a change in the context of a word mirrors a change in its meaning or usage. These models have undergone a surge in interest these last two years with the publication of several literature review articles [Tahmasebi *et al.*, 2018].

These models usually consist in dividing a temporal corpus into several time slices. The two most broadly used methods are *incremental updating* [Kim *et al.*, 2014] and *vector space alignment* [Hamilton *et al.*, 2016]. In the first one, an embedding matrix is trained on the first time slice of the corpus and updated at each successive time slice using the previous matrix as initialisation. For the second method, an embedding matrix is trained on each time slice independently. Due to the stochastic aspect of word embeddings, the vector space for each time slice is different: an alignment has to be performed by optimising a geometric transformation.The alignment method was proved to be superior to the incremental updating method, on a set of synthetic semantic drifts [Shoemark *et al.*, 2019]. It has been extensively used in the literature. However, these methods do not take into account the polysemy of words, summarising all the possibles senses into one vector at each time step. An exception is the system from Frermann and Lapata [2016] which analyses the evolution of sets of senses using a Bayesian model.

In parallel, the analysis of synchronic variations is mostly done through domain-specific word sense disambiguation (WSD). Some research use similarity measures between static word embeddings to analyse the variations in a vocabulary among several communities [Tredici and Fernández, 2017]. More recently, Schlechtweg *et al.* [2019] analyse both diachronic and synchronic drifts using static word embeddings with vector space alignment.

The recent rise of contextualised embeddings (for example BERT [Devlin *et al.*, 2019] or ELMO [Peters *et al.*, 2018]) brought huge change to the field of word representation. These models allow each token – each occurrence of a word – to have a vector representation that depends on its context. When pre-trained on large datasets, they improve the state-of-the-art on numerous NLP tasks. Similarly, contextualised embeddings can be used for better semantic change detection.

It was first used in a supervised way [Hu *et al.*, 2019]: for a set of polysemic words, a representation for each of their sense is learned using BERT. Then a pre-trained BERT model is applied to a diachronic corpus, extracting token embeddings and matching them to their closest sense embedding. Finally, the proportions of every sense is computed at each successive time slice, revealing the evolution of the distribution of senses for a target word. However, this method requires to know the set of senses of all target words beforehand. Another possibility is to use clustering on all token representations of a word, to automatically extract its set of senses [Giulianelli *et al.*, 2019; Martinc *et al.*, 2020a]. Our analysis in this paper derives from this last set of methods.

3 Model and Pipeline

We briefly describe the model BERT [Devlin *et al.*, 2019] and present the pipeline of detection and interpretation of word use variation.

3.1 Contextualised Embeddings Model: BERT

BERT stands for Bidirectional Encoder Representations from Transformers. It is a method for pre-training language representations that gets state-of-the-art results in a large variety of NLP tasks. The main idea relies on the principle of transfer learning: pre-training a neural network model on a known task with a substantial amount of data before fine-tuning it on a new task.

The architecture of BERT is a multi-layer bidirectional Transformer encoder [Vaswani *et al.*, 2017], a recent and popular attention model, applied to language modelling. The key element to this architecture is the bidirectional training, which differs from previous approaches. It is enabled by a new training strategy, Masked Language Model: 15% of the tokens in each input sequence are selected, from which 80% are replaced with a [MASK] token. The model is trained to predict the original value of these selected tokens using the rest of the sequence. A second training strategy is used, named Next Sentence Prediction (NSP): a set of pairs of sentences is generated for input, with 50% beings pairs of successive sentences extracted from a document, and 50% being two random sentences from the corpus. The model is trained to predict if the two sentences are successive or not.

BERT is mostly used in the literature as a pre-trained model before being fine-tuned on the task of interest, by

adding a task-specific layer to the architecture (Sentiment Classification, Named Entity Recognition...). On the contrary, what we are interested in when using BERT is the pre-trained language understanding model which, applied to any sequence, allows to extract contextualised representations for each token (feature-based approach).

3.2 Detecting Variations

We consider a corpus where each sequence is labelled with the time it was written, and the person who wrote it. The author can be characterised by several dimension (the community he belongs to, its geographical location...) that are the synchronic dimensions for the analysis.

We apply a pre-trained BERT model on this corpus; To get a vector representation of all tokens of a sequence, we concatenate the top four hidden layers of the pre-trained model, as advised by Devlin *et al.*[2019]. Thus, we obtain a vector representation for each token of each sentence.

In order to identify the various types of usages of a word, we want to apply a clustering algorithm to the set of token embeddings. Previous works using BERT rely on hand-picking a small amount of target words for semantic change analysis [Giulianelli *et al.*, 2019; Hu *et al.*, 2019; Martinc *et al.*, 2020b]. Our goal is to detect in the full vocabulary which words undergo a variation of usage; however, the clustering step is computationally heavy and can not be computed for a large vocabulary. Thus, we use a preliminary step to detect high-variation words, by extending the approach of [Martinc *et al.*, 2020a] to synchronic analysis.

For a given target word, we compute a variation metric for each of its dimensions of variation. First, we calculate the average token embedding on the full corpus, and the average token embedding for each class of the dimension (for example, for each author or for each year). Then, we take the mean of the cosine distance between each average class embedding and the full corpus average embedding.

We sort the vocabulary according to the variation measures, and select a limited list of target words from the top ranking words. For each selected word, we apply a clustering on all its token representations across the full corpus.

3.3 Clustering Token Embeddings

We use two clustering methods: K-Means and affinity propagation. The affinity propagation algorithm, less common than K-Means, is chosen for two reasons:

First, it has proven its efficiency in the literature for word sense induction [Alagić *et al.*, 2018], a task very close to what we want to achieve.

Second, it does not require the number of clusters to be selected manually, which is convenient for our task where the number of usages varies a lot depending on the word and is tricky to determine. Indeed, this number does not necessarily match the number of senses of the target word. As BERT does not induce perfectly semantic representations, the contextualised representations are heavily influenced by syntax [Coenen *et al.*, 2019]. Thus, the clusters obtained from the representations of a word do not naturally reflect the different senses of the word; more widely, it only reflects the different ways it is used.

Affinity propagation is an iterative clustering algorithm. The main idea is that all data points communicate by sending messages about their relative attractiveness and availability, using the opposite of the euclidean distance as similarity measure. Eventually, clusters of similar points emerge.

This algorithm often leads to a high number of clusters. This allows a very precise distinction of the different types of contexts the words appears in; however, in such situation with a high number of small clusters, it is much harder for a financial analyst to provide an interpretation of the different clusters and of the variation of word usage.

3.4 Analysing Clustering Results

After the clustering, all the occurrences of a word are distributed into clusters. Each token is labelled by its diachronic dimension (the time slice where the token appears), and its synchronic dimension (the class of the document).

We construct the probability distributions of the types of usages of a target word for each class of a dimension. For example, in the case of the time dimension, each token is associated with one time slice of our corpus. For each time slice, we extract the distribution of usages of the word across the clusters. We normalise it by the number of tokens. We obtain the probability distributions of clusters through the time dimension. The process is the same in the synchronic case.

We can compare these distributions together to extract several pieces of information:

1. How much the distributions of usages vary for the word through the dimension?
2. At what time a usage drift happens (for the diachronic dimension); which actor has a different usage distribution compared to the other ones?
3. What is the change about, which usages of the word are involved? How to make an interpretation of this change?

For the first element we use the Jensen-Shannon divergence (JSD), a metric to compare two probability distributions, and its generalisation to n probability distributions $d_1, d_2, \ldots, d_n$ [Ré and Azad, 2014]. With H being the entropy function, the generalised JSD is:

$$\mathrm{JSD}(d_1, d_2, \ldots, d_n) = H\left(\frac{\sum_{i=1}^{n} d_i}{n}\right) - \frac{\sum_{i=1}^{n} H(d_i)}{n} \quad (1)$$

It is applicable in both synchronic and diachronic cases.

For the second element, we compare each distribution with the average distribution of the full dimension. For example, in the diachronic case, we average the distributions for all the time slices element-wise. Then, we compute the Jensen-Shannon divergence with the global average distribution.

In order to capture the clusters involved in the variation, we identify the ones that have an uneven distribution across all the elements of the dimension. It allows for example to find the clusters specific to a given actor, the clusters that vary the most, or the ones that appear or disappear through time.

Finally for the third element, once the clusters of interest are identified, we can get an interpretation of the usages associated with them using two methods. One the one hand, we identify the centroids of the clusters: the example (in our

case, the sentence) that is the closest to the centroid is assumed to be representative of the context of the tokens inside the cluster. Thus, we observe these central sentences to get a preliminary idea of the word usages in context. On the other hand, we set up a keyword detection method to characterise the different clusters in relation to one another. Relying on the tf-idf (Term Frequency - Inverse Document Frequency) principle, each cluster containing a set of sentences, we consider them as documents and the set of clusters as a corpus. The goal is to identify the most discriminant words for each cluster. The stop-words and the words appearing in more than 50% of the clusters are excluded from the analysis. We compute the tf-idf score of each word in each cluster. The words with the highest score in a cluster are the most important for the analysis of this cluster: they are used as keywords to ease its interpretation.

4 Experiments

We apply the word usage variation detection pipeline to two financial corpora across several dimensions in addition to time. For our experiments, we use the English BERT-base-uncased model from the library `Transformers`[1] with 12 attention layers, an output layer of size 768 and 110M parameters.

4.1 Data

We use two financial corpora spanning two decades: a corpus of annual financial reports (10-K) of U.S. companies extracted from the Securities Exchange Commission database (SEC-Edgar), and a corpus of central bank statements.

The SEC-Edgar filings were extensively studied in the literature. From the diachronic point of view, Purver *et al.* [2018] extract subsets of the annual reports of 30 companies from the Dow Jones Industrial Average (DJIA) from 1996 to 2015. They manually select a set of 12 financial terms and investigate changes in lexical associations, by looking at the evolution of the similarity between pairs of two terms. More recently, [Desola *et al.*, 2019] fined-tune BERT separately on two corpora of SEC-EDGAR filings (from years 1998-1999 and years 2017-2019). For three selected words (*cloud*, *taxes* and *rates*), they compare the embeddings from the two periods using cosine similarity. None of these works are fully unsupervised.

We scrape[2] the SEC-EDGAR reports[3] from the 500 biggest companies in the US, between 1998 and today. Similarly to [Purver *et al.*, 2018], we extract the Part I and the Items 7 and 7A from the Part II of the 10-K annual reports. These sections mainly describe the activity of the company and its operations and management. We exclude the year 2019 from the analysis, as many documents of that year are not available yet. We end up with 8676 documents spanning 20 years. It amounts to a total of 7.3 million sentences.

This corpus is very rich for synchronic analysis. Each document is written by one company, and for each company, we extract additional data: its stock exchange (NYSE, NASDAQ, OTC) and its Standard Industrial Classification[4] (SIC) code. The latter indicates the business line of the company; the classification is divided into 7 Offices and sub-divided into 444 Industries. Thus, we can detect drifts across several dimensions, from the most to the least fine-grained: by company, by Industry, by Office, and by Stock Exchange.

The second corpus assembles all the official statements of two central banks, the European Central Bank (ECB) and the US Federal Reserve Bank (Fed) from June 1998 to June 2019[5]. These statements report the economic situation and expose the policy decisions of the central banks. This corpus was previously studied through sentiment analysis [Buechel *et al.*, 2019]. It is composed of 230 documents from the ECB and 181 from the Fed, and contain a total of 14604 sentences; it is heavily unbalanced towards the ECB (more than 75% of sentences), as the Fed statements are usually shorter.

Both corpora are divided into 20 yearly time steps. Stopwords are removed and we build the vocabulary with all words having at least 100 occurrences in the corpus.

	Time	**Source**
1	households	measures
2	labor	committee
3	holdings	rate
4	securities	employment
5	accomodative	developments
6	sectors	support
7	monetary	pressures
8	housing	price
9	sales	stability
10	loan	market

Table 1: Top 10 words with highest variation measure (from section 3.2) for the time dimension and the source dimension on the Central Bank Statements corpus

4.2 Selecting Target Words

For both corpora, we conduct the preliminary step on the full vocabulary.

On the SEC-Edgar corpus, the frequency of some words is very high (for example the word *million* appears 1.4 million times). To speed up the process, we sample 3000 sentences for each word. We extract the embedding of the target word using BERT. Then, we compute the variation measures from section 3.2 by year, by company, by Industry, by Office, and by Stock Exchange. We do the same in the Central Banks Statements corpus, by year and by source.

As an example, the words with highest variation for the time dimension and the source dimension on the Central Bank Statements corpus are showed in Table 1. For the source dimension, we keep only the words with a threshold of presence of at least 50 occurrences per source. Words such as *labor* are absent from the FED statements because of orthographic divergence between UK English and US English.

[1] Available at https://huggingface.co/transformers/

[2] Using https://github.com/alions7000/SEC-EDGAR-text

[3] Extracted from https://www.sec.gov/edgar.shtml

[4] Described in https://www.sec.gov/info/edgar/siccodes.htm

[5] We thank Sven Buechel from Jena University Language & Information Engineering (JULIE) Lab for sharing the corpus with us.

Method	S-score	JSD-synchronic	JSD-diachronic
Aff-prop	0.267	0.829	2.519
KMeans3	0.213	0.342	0.523
KMeans5	0.215	0.467	0.856
KMeans7	0.218	0.537	1.088

Table 2: The average values of silhouette score, JSD by source and JSD by year for all the target words

Source:	KM2	KM3	KM5	KM7	KM10
	0.389	0.484	0.579	0.596	0.630
Time:	KM2	KM3	KM5	KM7	KM10
	0.325	0.328	0.289	0.282	0.282

Table 3: For both dimensions, the correlation between the JSD from affinity propagation clustering and the JSD from K-Means (KM) with different k.

For each dimension, we select the 10% words with highest variation measure as target words for the clustering step.

4.3 Comparison of the Clustering Algorithms

We apply both K-Means and affinity propagation on the set of token embeddings of each target word. In the case of K-Means, for each word we try different values of the number of clusters k ranging in $[2:10]$. To evaluate the quality of a clustering, we compute its silhouette score for each target word. Then, we extract the probability distributions across each dimension (for example the distribution of each year for the time dimension). We apply the generalised Jensen-Shannon Divergence (JSD) on the set of probability distributions to measure the level of usage variation of the word.

We focus on the Central Bank Statements Corpus to analyse the results of the clustering. The average values of silhouette score, JSD by source and JSD by year for all target words of this corpus for different algorithms are in Table 2. It should be recalled that the silhouette score takes values between 0 and 1, a value close to zero signalling a low clustering quality. Plus, while the JSD between two distributions takes values between 0 and 1, the generalised version to n dimensions is restricted by $\log_2(n)$. For example for the temporal dimension in the Central Bank Statements corpus, the 20-years period leads to an upper bound being equal to $\log_2(20) \approx 4.32$.

According to Table 2, the average silhouette score is the highest for the affinity propagation algorithm. Moreover, the average JSD for both dimensions increases with the number of clusters for the algorithm K-Means. The correlation between the number of clusters (k from 2 to 10) and the average JSD at each k, is high and positive (0.962 for the JSD by source and 0.986 for the JSD by year). We also inspect the number of clusters for the affinity propagation algorithm. It ranges from 4 to 450, with an average number of 61 clusters. However, the correlation between the number of clusters and the JSD is not significant. On the contrary, the correlation between the number of clusters of the affinity propagation algorithm and the silhouette score is -0.29: words with very diversified usages are associated with a clustering of lower quality. Finally, one can evaluate the accordance between the affinity propagation and K-Means algorithms by computing the correlation between their respective JSD for all words. According to Table 3, the correlation between affinity propagation JSD and K-Means JSD increases with k for the Source dimension, while it is relatively stable for the Time dimension.

Label	**Description**	**%**
0	Office of Energy & Transportation	15.1
1	Office of Finance	12.5
2	Office of Life Sciences	14.7
3	Office of Manufacturing	19.7
4	Office of Real Estate & Construction	8.2
5	Office of Technology	13.1
6	Office of Trade & Services	16.7

Table 4: Label and proportion of business line with SIC classification in the SEC-Edgar corpus

4.4 Interpreting the Clusters

We focus on the SEC-Edgar Corpus for this last step. We present one example for the diachronic dimension and one for the synchronic dimension, in order to show the different possibilities in terms of interpretation.

For the synchronic dimension, we study the distribution of usages of the word *client* by Office (business line). It is one of the words with the highest JSD for this dimension. The silhouette score is the highest using K-Means algorithm with $k = 4$. All the Offices are listed in Table 4; The normalised distributions of clusters for each of them are in the left part of Figure 1. We apply our interpretation pipeline to identify the clusters that have an uneven distribution, and the Offices that are involved. Using the keyword extraction method, we select the most representative words for each cluster (Table 5, left). The cluster 1 is the most unevenly distributed, and appears mostly in documents belonging to the Real Estate & Construction Office. The keywords associated with this cluster involve the idea of paying (*cost*, *fees*) and negativity (*risk*, *loss*). On the contrary, the clusters 2 and 3 are relatively similarly allocated in the different Offices. Their keywords correspond to the classical definition of a client in a company. Finally, the cluster 0 is characterised by vocabulary from the semantic field of digital technologies (*server*, *applications*...): the clustering algorithm was able to identify this specific meaning of the target word.

For the diachronic dimension, we study the distribution of usages of the word *crisis* by year (Figure 1, right). The highest silhouette score corresponds to the K-Means algorithm with $k = 5$. The keywords for these 5 clusters can be found in Table 5 (right side). We can identify clear temporal tendencies in the figure. The proportions of the clusters 0 and 4 are decreasing through time, while the clusters 1 and 2 are growing. The extraction of keywords allows to differentiate the 5 usages of the word *crisis*. For example, the cluster 1 is associated with vocabulary of the domain of marketing and media. It is almost non-existent before the year 2004, and

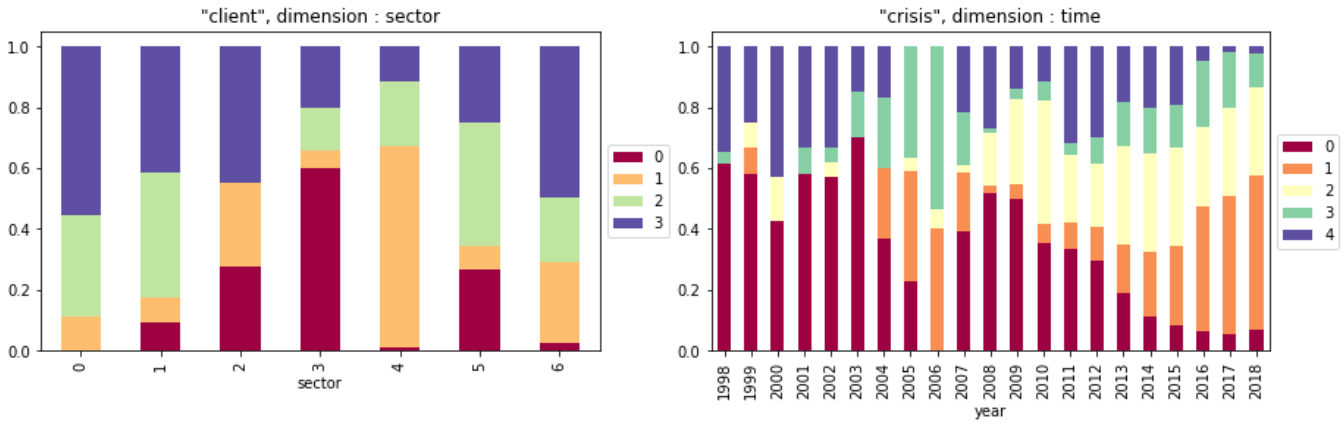

Figure 1: Distribution of clusters per Office for the word *client* (left) and per year for the word *crisis* (right) in the SEC-Edgar corpus. The Offices are described in Table 4

N^o	Keyword examples - Word = *client*
0	server, products, data, applications, services, systems
1	revenue, contract, risk, costs, loss, business, fees
2	assets, funds, cash, interest, balances, investment
3	services, business, revenue, growth, management, products

N^o	Keyword examples - Word = *crisis*
0	liquidity, funding, contingency, cash, collateral, outflows
1	marketing, business, management, design, advertising, media
2	european, debt, credit, sovereign, countries, eurozone, banks
3	financial, accident, capital, regulatory, loss, liquidity, funding
4	credit, financial, global, markets, debt, european, recession

Table 5: List of clusters and keyword examples for the words *client* (left) and *crisis* (right) in the SEC-Edgar Corpus

is rapidly growing. The cluster 2 is related to the crisis of the debt of the European countries; it appears and grows after 2008. The cluster 3 can be found across all the period; it is associated with slightly negative words (*accident* and *loss*), similarly to the cluster 4 (associated with *debt* and *recession*) whose proportion decreases since 2010.

However, one has to be wary of the selection of the number of clusters using the silhouette score. Sometimes, it leads to choose a low amount of clusters that may hide some valuable information. For example, for the target word *insurance*, the silhouette score is maximum for K-means with $k = 2$. However, using $k > 5$, a cluster appears that belong mostly to sector 4 (Office of Real Estate Construction); it is associated with the keywords *property* and *investment*, showing a new aspect of the concept of insurance specific to this sector.

Overall, the disparities in vocabulary and connotation between clusters are encouraging. The clustering allows to identify variations in meaning as well as usage. In particular, the ability to detect clear temporal tendencies in the cluster distributions could allow a financial analyst to link these clusters with real-world events, and have a deeper understanding of the phenomenons behind them.

5 Discussion

In this paper, we investigate the ability of the contextualised embeddings model BERT to detect meaningful synchronic and diachronic word use change in a financial corpus. We showed that using contextualised embeddings associated with clustering allows to automatically detect variations in the use of a word across any dimension. However, even though the keyword extraction method allows to gain insight on the interpretation of the clusters, it still requires some domain-specific knowledge. A crucial next step is to build on this pipeline to propose an evaluation method.

To this end, we would like to link the detected word usage variations with numerical indicators. On the one hand, it would offer a better understanding of the implications of the variations of word usage and complement their interpretations. On the other hand, it would allow a form of evaluation of our method. For example, we can analyse the correlation between the cluster distributions of the token embeddings of the word *unemployment* by Office in the SEC-Edgar reports with the real unemployment curve by Office on the same time period.

To fully leverage the ability of this pipeline to detect and to interpret word usage variations, our method can straightforwardly be extended in a streaming way. Any new document can be included in the analysis, be it a new central bank statement, company report, or in a classical streaming data situation such as daily financial news or tweets. The new document has to be tokenised and the contextualised embeddings extracted; then, the clustering can be updated using incremental clustering methods. For example, several incremental affinity propagation algorithms, adapted to streaming data, are proposed in the literature [Ajithkumar and Wilson, 2017]. The new token embeddings can either be added to an existing cluster, thus modifying the distribution, or creating a new cluster.

References

[Ajithkumar and Wilson, 2017] S. Ajithkumar and Praveen K Wilson. A survey paper on clustering data using incremental affinity propagation. In *IOSR Journal of Computer Engineering (IOSR-JCE)*, 2017.

[Alagić *et al.*, 2018] Domagoj Alagić, Jan Šnajder, and Sebastian Padó. Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Buechel *et al.*, 2019] Sven Buechel, Simon Junker, Thore Schlaak, Claus Michelsen, and Udo Hahn. A time series analysis of emotional loading in central bank statements. In *Proceedings of the Second EcoNLP Workshop*, pages 16–21, Hong Kong, November 2019.

[Coenen *et al.*, 2019] Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. Visualizing and measuring the geometry of bert. In *NeurIPS*, 2019.

[Desola *et al.*, 2019] Vinicio Desola, Kevin Hanna, and Pri Nonis. Finbert: pre-trained model on sec filings for financial natural language tasks. 08 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[Frermann and Lapata, 2016] Lea Frermann and Mirella Lapata. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45, 2016.

[Giulianelli *et al.*, 2019] Mario Giulianelli, Raquel Fernandez, and Marco Del Tredici. Contextualised word representations for lexical semantic change analysis. In *EurNLP*, 2019.

[Gulordava and Baroni, 2011] Kristina Gulordava and Marco Baroni. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop*, pages 67–71, 2011.

[Hamilton *et al.*, 2016] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, 2016.

[Hu *et al.*, 2019] Renfen Hu, Shen Li, and Shichen Liang. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy, July 2019.

[Kim *et al.*, 2014] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, 2014.

[Martinc *et al.*, 2020a] Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 343–349, New York, NY, USA, 2020. Association for Computing Machinery.

[Martinc *et al.*, 2020b] Matej Martinc, Petra Kralj Novak, and Senja Pollak. Leveraging contextual embeddings for detecting diachronic semantic shift. In *LREC*, 2020.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[Peters *et al.*, 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *ArXiv*, abs/1802.05365, 2018.

[Purver *et al.*, 2018] Matthew Purver, Aljosa Valentincic, Marko Pahor, and Senja Pollak. Diachronic lexical changes in company reports : An initial investigation. In *Proceddings of the First Financial Narrative Processing Workshop (FNP 2018)*, 2018.

[Ré and Azad, 2014] Ma Ré and Rajeev Azad. Generalization of entropy based divergence measures for symbolic sequence analysis. *PloS one*, 9:e93532, 04 2014.

[Schlechtweg *et al.*, 2019] Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy, 2019.

[Shoemark *et al.*, 2019] Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 EMNLP-IJCNLP Conference*, pages 66–76, Hong Kong, China, 2019.

[Tahmasebi *et al.*, 2018] Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to diachronic conceptual change. *CoRR*, 1811.06278, 2018.

[Tredici and Fernández, 2017] Marco Del Tredici and Raquel Fernández. Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics*, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, 2017.

[Wiedemann *et al.*, 2019] Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of KONVENS 2019*, Erlangen, Germany, 2019.