Applied Intelligence
https://doi.org/10.1007/s10489-023-04452-4

Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with Latent Dirichlet Allocation

Silvia García-Méndez[1] · Francisco de Arriba-Pérez[1] · Ana Barros-Vila[1] · Francisco J. González-Castaño[1] · Enrique Costa-Montenegro[1]

Accepted: 1 January 2023

Abstract

Financial news items are unstructured sources of information that can be mined to extract knowledge for market screening applications. They are typically written by market experts who describe stock market events within the context of social, economic and political change. Manual extraction of relevant information from the continuous stream of finance-related news is cumbersome and beyond the skills of many investors, who, at most, can follow a few sources and authors. Accordingly, we focus on the analysis of financial news to identify relevant text and, within that text, forecasts and predictions. We propose a novel Natural Language Processing (NLP) system to assist investors in the detection of relevant financial events in unstructured textual sources by considering both relevance and temporality at the discursive level. Firstly, we segment the text to group together closely related text. Secondly, we apply co-reference resolution to discover internal dependencies within segments. Finally, we perform relevant topic modelling with Latent Dirichlet Allocation (LDA) to separate relevant from less relevant text and then analyse the relevant text using a Machine Learning-oriented temporal approach to identify predictions and speculative statements. Our solution outperformed a rule-based baseline system. We created an experimental data set composed of 2,158 financial news items that were manually labelled by NLP researchers to evaluate our solution. Inter-agreement Alpha-reliability and accuracy values, and ROUGE-L results endorse its potential as a valuable tool for busy investors. The ROUGE-L values for the identification of relevant text and predictions/forecasts were 0.662 and 0.982, respectively. To our knowledge, this is the first work to jointly consider relevance and temporality at the discursive level. It contributes to the transfer of human associative discourse capabilities to expert systems through the combination of multi-paragraph topic segmentation and co-reference resolution to separate author expression patterns, topic modelling with LDA to detect relevant text, and discursive temporality analysis to identify forecasts and predictions within this text. Our solution may have compelling applications in the financial field, including the possibility of extracting relevant statements on investment strategies to analyse authors' reputations.

Keywords Natural language processing · Knowledge extraction · Latent Dirichlet Allocation · Personal finance management · Financial news analysis · Temporality analysis

1 Introduction

1.1 Motivation

New efficient algorithms [1–4] and the prolific sources of online information have boosted applied data analysis research. In this scenario, Natural Language Processing (NLP) techniques are being successfully applied to unstructured textual data [5–8], from the simplest approaches that use morphological information as input [9] to more complex methodologies that take advantage of syntactic patterns and semantic relations [10].

Silvia García-Méndez, Francisco de Arriba-Pérez, Ana Barros-Vila, Francisco J. González-Castaño and Enrique Costa-Montenegro are contributed equally to this work.

✉ Silvia García-Méndez
sgarcia@gti.uvigo.es

Extended author information available on the last page of the article.

Published online: 10 March 2023

Financial Knowledge Extraction (KE) is of particular interest. NLP techniques have been used to apply a wealth of market forecasting research to financial news, economic reports and financial expert comments [11]. Financial news describes relevant market events, their causes and their possible effects. Transferring human associative discourse capabilities [12] from this type of content is challenging.

1.2 Financial knowledge extraction

Some representative examples of financial KE include information extraction from financial news for firm-based monitoring [13]; analysis of financial risk such as volatility [14] and Personal Finance Management applications [15], among other interesting use cases [7]. Most of these KE systems engineer specific features of the content with their target in mind [16].

It is well known that there is a strong relation between mass media news and stock market state [17, 18]. Previous research has shown that information published in media outlets or shared financial data in printed media, radio, television, and web sites is correlated with future stock market events [19]. Apart from providing valuable objective information in financial news, authors speculate about market events within political, social and cultural contexts. In these unstructured texts the discourse flows around certain key statements and predictions, and an automatic financial news analysis system should distinguish between less relevant data and predictions to gather knowledge to assist investors in decision making [20].

1.3 Temporality at the discursive level

Temporal representation in texts and speculative statements in particular is based on semantic combinations of certain linguistic structures and elements [21]. However, the vast majority of works on temporality research at the discursive level have simply focused on verb tenses [22], ignoring their semantic context.

1.4 Research goal and main contribution

Our research is a case of financial News Analysis (NA) [13–15] within the field of Intelligence Amplification, which has lately gained attention [23–25] as a means of enhancing the understanding and reasoning capabilities of automatic KE solutions and transferring human associative discourse capabilities to expert systems. Our case contributes to solving the problem of extracting relevant text from financial news and, within that relevant text, identifying forecasts and predictions. Our solution may be valuable in helping inexpert stockholders to process more financial news more efficiently.

To the best of our knowledge, this is the first study to propose an approach for the automatic detection of relevant events in financial NA based on the joint consideration of relevance and temporality analysis at the discourse level.

1.5 Approach

Our approach comprises:

- Multi-paragraph topic segmentation and co-reference resolution to separate author expression patterns.
- Detection of relevant text through topic modelling with Latent Dirichlet Allocation (LDA), outperforming a rule-based system.
- Identification of forecasts and predictions within relevant text using discursive temporality analysis and Machine Learning (ML).

We demonstrate the performance of these features using an experimental data set composed of news items from widely used financial sources. The final data set had 2,158 financial news items similar in size to or even larger than other studies in the literature [26–31] that were manually labelled by NLP researchers.

1.6 Structure of the paper

The rest of this article is organised as follows. Section 2 reviews related work on KE and NA solutions. Section 3 describes our automatic system for detecting relevant financial events based on NLP and ML techniques. Section 4 presents the text corpus and numerical evaluation of our solution. Finally, Section 5 concludes the article.

2 Related work

Stock market research is based on fundamental and technical approaches [32]. Fundamental approaches involve performing stock market forecasts using numerical data such as price variation. Technical approaches, in turn, focus on the temporal dimension of financial events. They apply trend modelling techniques to historical asset data forecasts.

Previous research on Data Mining and KE for stock market screening on textual data has considered financial news [14, 33], stockholder comments in blogs [18] and social networks [34]. These systems apply NLP techniques [11] or ML models [2, 14], which may be supervised [35], relying on automatic or manually annotated data sets, or unsupervised [36], taking into account the peculiarities of input data and descriptive patterns. The simplest approach consists of using a vector representation of the content and weighting the terms once meaningless elements, such as prepositions [37], are removed. More complex approaches, like the one presented by De Arriba-Pérez et al. (2020) [38], seek to identify syntactic and semantic patterns as key descriptors of financial news through lexica, grammar and name entity recognition techniques.

Traditional extraction methods for filtering relevant text in this context comprise manual[1] and automatic[2] pattern discovery approaches [40]. The former require large knowledge bases, such as dictionaries and lexica, and rule sets. They tend to be constrained by specific application domains. Automatic approaches include simple statistical and more demanding, complex linguistic approaches, in addition to the previously mentioned ML solutions [39]. TF-IDF [41] is remarkably simple, but it has been reported to under-perform on professional texts, as in our case. Alternative solutions combine the previous techniques with knowledge heuristics such as position, length and text format information. A more competitive solution is fuzzy logic for sentence scoring. However, this lacks adaptability and requires manual rule generation, which directly affects performance [42].

Unsupervised extraction is highly practical because it eliminates the burden of text tagging. Nevertheless, many KE solutions rely on supervised methodologies. Examples those of Gottipati S et al. (2018) [43], who designed an ML course improvement solution based on student feedback and compared its performance to a rule-based method; López-Úbeda et al. (2021) [44], who extracted relevant information from radiological reports; and Verneer et al. (2019) [45], who proposed a relevance detection system from social media messages (although they noted the great potential of LDA as an alternative).

Among extraction solutions developed to detect relevant topics from news pieces (setting aside temporality analysis), Jacobs et al. (2018) [46] developed a supervised model for economic event extraction in English news using a sentence-level classification approach, as in our case; Oncharoen et al. (2018) [47] applied the Open Information Extraction system to represent the news data as tuples (actor, action and object); Carta et al. (2021) [48] employed a real-time domain-specific clustering-based approach for event extraction in news; and Harb et al. (2008) [49] presented a linguistic-based opinion extraction system for blogs.

Assuming there is a direct causal relationship between financial news and asset prices [17], some authors have explored both ML and other sophisticated techniques such as deep learning to gather context-dependent information for stock market screening [50]. Worthy of note in this respect is the Naive Bayes model by Atkins et al. (2018) [14] for predicting stock market volatility, which employed as input word-topic correspondence feature vectors obtained with LDA. Unlike our proposal, this model considered news content as a whole and did not differentiate non-relevant from relevant parts for their target application. Shilpa & Shambhavi (2021) [50] presented a prediction framework based on sentiment analysis and stock market technical-indicator features. Temporality was not considered.

Prior work has addressed linguistic [51], template-based [52] and statistical news summarisation approaches [53]. State-of-the-art summarisation systems may be extractive [53] or abstractive [54]. Extractive summaries, which are more akin to our goal, extract key sentences directly from the input text. These sentences are ranked by importance and selected if they pass a threshold. Query-focused and update summarisation approaches [55] also deserve consideration as they retrieve information tailored to a specific audience. The summarisation of online financial news in our work focuses on financial investors. The temporal dimension, expressed as discursive temporality, is crucial to us because relevant text in finance-related news may include in addition to factual information, speculations or predictions, whether quantitative or not. In further relation to summarisation, template-based systems on financial NA [56] were limited in early research due to their computational load, the laborious task of defining the templates and their lack of flexibility.

KE solutions, and in particular, financial NA systems, have not paid sufficient attention to temporal analysis. The vast majority simply use temporal references provided by timestamps or verb tenses. Evers-Vermeul et al. (2017) [22], for example, simply noted that as linguistic markers, verb tense suffixes express temporal order and coherence relations through text. Our work goes a step further by analysing sentence-level temporality through syntax and semantics, and detecting temporal elements, expressions and the patterns in which they are arranged.

Summing up, Table 1 compares the most relevant work related to our proposal. Our main contribution is the detection of relevant statements on financial news including forecasts and predictions. To do this, our system automatically groups related data and filters out background information. In brief, we present a novel technique combining LDA analysis of automatically segmented news with temporality analysis at the discourse level. To our knowledge, this is the first NA approach that jointly considers relevance and temporality at the discourse level.

3 System architecture

In this section, we describe our system for the automatic detection of relevant financial events using NLP techniques

[1] *E.g.*, cascaded and non-deterministic finite state automatons, semantic information extraction solutions, etc.

[2] *E.g.*, supervised and unsupervised ML methods, being the first more extended [39].

Table 1 Comparison with related works

Authorship	Source	Relevance analysis	Temporality analysis	Technique
Atkins A et al. (2018) [14]	News	✗	Time series data only	LDA and ML
Oncharoen P et al. (2018) [47]	News	✗	Temporal relations only	CNN and LSTM
De Arriba-Pérez F et al. (2020) [38]	Micro-blogging data	✗	✗	NLP and ML
Carta S et al. (2021) [48]	News and micro-blogging data	✓	✗	Clustering
Shilpa B et al. (2021) [50]	News	✗	✗	NLP and ML
Our proposal	News	✓	✓	Multi-paragraph topic segmentation, co-reference, LDA, discursive analysis and ML

and KE algorithms. Figure 1 shows the scheme of the system. First, we segment the input text to group together closely related information. Then, we apply co-reference resolution to discover internal dependencies in the news content among key references to assets. The next stage is the tag processing stage, which consists of the detection, homogenisation and replacement of financial terms. This is followed by relevant topic modelling and temporal analysis to identify predictions and speculative statements, and ultimately provide investors with a synthesised version of financial news highlighting pertinent information, such as asset performance and forecasts/predictions as a summary.

3.1 Multi-paragraph topic segmentation

The first stage of the multi-paragraph topic segmentation process applies the enhanced version of the TextTiling algorithm [57] to segment news content into subtopic paragraphs.

TextTiling exploits lexical co-occurrence and discourse distribution patterns to identify subtopic paragraphs within a text with reasonable precision from a human's perspective. The algorithm compares, in sequence, the similarity of adjacent text divisions of similar length. If the vocabulary in the first and second parts of the comparison differ, the division is considered a split point. We set a minimum text length of 500 characters to apply the algorithm. Otherwise, no segmentation is performed.

The rationale behind this first stage is the assumption that text segments must be coherent, self-contained information units. Explicit paragraphs of financial news are less useful in our case because, quite often, they just break up the text layout to facilitate readability. It is the informal, inner structure based on key statements and predictions about certain assets or stock markets that interests us.

Table 2 shows an example of the original content of a financial news item. Table 3 shows the same item after segmentation with TextTiling (for conciseness, we present just the first two segments). In this example, the first paragraph refers to the current state of the asset and the second describes its past performance. Logically, TextTiling does not always guarantee such a degree of coherence, but in our case, it contributes to the overall efficiency of our system as a building block.

3.2 Co-reference resolution

The purpose of co-reference resolution is to replace references with meaningful words, which improves the performance of the subsequent LDA stage. Specifically, after segmenting the text, we use the Neural Network (NN) by Clark et al. (2016) [58] to generate high-dimensional

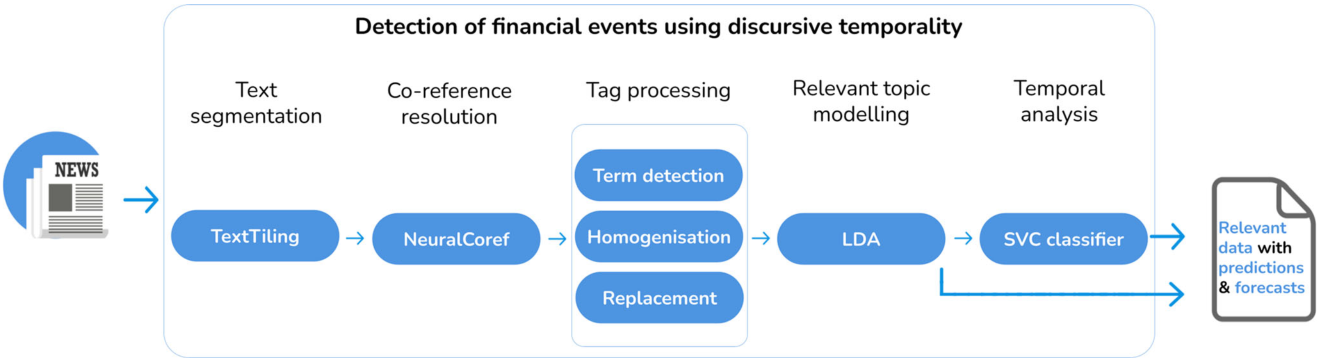


Fig. 1 Proposed system

Table 2 Example of news entry

Verizon Communications (NYSE:VZ) is proving to be a stable but undervalued company in the market today. In fact, VZ stock is worth at least 55% more than its price today using an analysis of its dividend yield, its own P/E ratio history, and a comparison with its peers. The company reported “boring” earnings, according to Barron’s magazine for Q2 on July 24. But the magazine says “boring is good” in this market. They point out that the communications business is not a bad place to be in a pandemic. For example, Verizon said that its earnings on a non-GAAP adjusted basis was $1.18 per share. This was only 3% lower than a year ago. On an adjusted EBITDA (earnings before interest, taxes, depreciation, and amortization) basis, its cash flow was down 5%. However, Verizon generated higher cash flow. For example, first-half 2020 cash flow from operations of $23.6 billion, an increase of $7.7 billion from first-half of 2019. This represents a huge increase of over 206%. Moreover, its free cash flow (FCF) in the first half was $13.7 billion, an increase of 74.1 percent year over year. But consider this. Verizon trades for a paltry 12.3 times this year’s expected earnings and just 12 times next year[...]

Table 3 Example of news entry after applying multi-paragraph TextTiling segmentation

ID	TextTiling segment
0	Verizon Communications (NYSE:VZ) is proving to be a stable but undervalued company in the market today. In fact, VZ stock is worth at least 55% more than its price today using an analysis of its dividend yield, its own P/E ratio history, and a comparison with its peers
1	The company reported “boring” earnings, according to Barron’s magazine for Q2 on July 24. But the magazine says “boring is good” in this market. They point out that the communications business is not a bad place to be in a pandemic. For example, Verizon said that its earnings on a non-GAAP adjusted basis was $1.18 per share[...]

Table 4 Example of news entry before and after applying co-reference resolution

Before co-reference resolution	After co-reference resolution
Verizon Communications (NYSE:VZ) is proving to be a stable but undervalued company in the market today. In fact, VZ stock is worth at least 55% more than **its** price today using an analysis of **its** dividend yield, **its** own P/E ratio history, and a comparison with **its** peers	Verizon Communications (NYSE:VZ) is proving to be a stable but undervalued company in the market today. In fact, VZ stock is worth at least 55% more than **VZ stock** price today using an analysis of **VZ stock** dividend yield, **VZ stock** own P/E ratio history, and a comparison with **VZ stock** peers

Table 5 Example of news content before and after applying financial term detection, homogenisation and replacement procedures

Before	After
Verizon Communications (NYSE:VZ) is proving to be a stable but undervalued company in the market today. In fact, **vz stock** is worth at least **55%** more than **VZ stock** price today using an analysis of **VZ stock** dividend yield, **VZ stock** own **P/E** ratio history, and a comparison with **VZ stock** peers	**TICKER (STOCK:TICKER_ABR)** is proving to be a stable but undervalued company in the market today. In fact, **TICKER_ABR** stock is worth at least **NUM** more than **TICKER_ABR stock** price today using an analysis of **TICKER_ABR stock** dividend yield, **TICKER_ABR stock** own **FIN_ABR** ratio history, and a comparison with **TICKER_ABR stock** peers

Table 6 Example of LDA analysis of relevant information

Text	LDA score
TICKER (**STOCK:TICKER_ABR**) is proving to be a stable but undervalued company in the market today	0.847
In fact, **TICKER_ABR** stock is worth at least **NUM** more than **TICKER_ABR** **stock** price today using an analysis of **TICKER_ABR** **stock** dividend yield, **TICKER_ABR** **stock** own **FIN_ABR** ratio history, and a comparison with **TICKER_ABR** **stock** peers	0.948
TICKER reported "boring" earnings, according to Barron's magazine for **FIN_ABR** on **DATE**	**0.571<0.8**
For example, **TICKER** said that **TICKER** earnings on a non-**FIN_ABR** adjusted basis was **NUM** per share	0.825
However, **TICKER** generated higher cash flow	0.870
Moreover, this free cash flow (**FIN_ABR**) in the first half was **NUM** billion, an increase of **NUM** percent year over year	0.897
TICKER trades for a paltry **NUM** times this year's expected earnings and just **NUM** times next year	0.934

vector representations for co-reference compatibility of cluster pairs. The NN is composed of two task-oriented sub-networks: a mention-pair encoder and a cluster-pair encoder, which create the distributed representations, and the cluster and mention-ranking models to score the pairs of clusters.

Table 4 provides an example of a news entry after this procedure. Note how implicit asset references have been replaced by explicit ones, which is essential for the next stages.

3.3 Tag processing: financial term detection, homogenisation and replacement

After text segmentation and co-reference resolution, the tag processing stage homogenises the input for the subsequent LDA stage. First, asset identifiers are detected using our financial lexica[3] on stock markets, tickers and currencies. In addition, we search for words such as *company*, *enterprise*, *manufacturer* and *shareholder*, which may refer to an asset. We then detect dependencies between these key terms. Next, we replace all references to stock markets, assets, asset abbreviations and currencies with the tags STOCK, TICKER, TICKER_ABR and CURRENCY, respectively. Using the same lexica, we also replace financial terms and abbreviations with the tag FIN_ABR.

A search is also made for capitalised proper names in the news items. Initially, these names are checked in the above and replaced by FIN_ABR, TICKER_ABR and TICKER tags, as appropriate, when there is a match. In the absence of a match, they are replaced by category tags taken from an entity recognition database.

We homogenise numerical values and dates, and group dates and times under the tag DATE and quantitative terms under the tag NUM using a Name Entity Recogniser (NER, see Section 4.1). Table 5 provides a complete example of co-reference resolution and tag processing with detection of financial terms, homogenisation and replacement using the tools specified in Section 4.1.

3.4 Relevant text detection with LDA topic modelling

LDA [59] is an unsupervised algorithm that identifies different topics in a particular document, but it can be generalised to unknown documents if they belong to the same domain and share similar context and structure [60]. We use it to differentiate between relevant and less relevant information in segments ("documents" in this section) produced from financial news content. Our goal is thus to discover relevant information in financial news by separating it from non-relevant information. Note that this type of news has a rather characteristic structure where relevant information is often presented along with precise contextual data and expression patterns, unlike other conventional news items.

Therefore, we employ a Dirichlet distribution with two topics (1). The training algorithm iterates to minimise the number of topics per word and document.

$$P(Z, W, \theta, \phi; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \times \prod_{i=1}^{2} P(\phi_i; \beta) \times \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\phi_{Z_{j,t}}) \quad (1)$$

- Z is the set of target topics, two in this work.
- W is the set of words (once stop-words[4] are discarded), with size N.

[3] Available at https://www.gti.uvigo.es/index.php/en/resources/14-resources-for-finance-knowledge-extraction , October 2022.

[4] Available at bit.ly/3yzvXqJ, October 2022.

Table 7 Temporal features in the ML temporal analysis model

Name	Description
{Pst,Prs,Fut}DepSub	Number of {past,present,future} tense verbs from the dependency analysis when the asset is the subject of the clause
GlobalDepSub	Global temporality by majority voting from the dependency analysis when the asset is the subject of the clause
{Pst,Prs,Fut}DepSubObj	Number of {past,present,future} tense verbs from the dependency analysis when the asset is the subject or object of the clause
GlobalDepSubObj	Global temporality by majority voting from the dependency analysis when the asset is the subject or object of the clause
{Pst,Prs,Fut}ProxSub	Number of {past,present,future} tense verbs from the proximity analysis when the asset is the subject of the clause
GlobalProxSub	Global temporality by majority voting from the proximity analysis when the asset is the subject of the clause
{Pst,Prs,Fut}ProxSubObj	Number of {past,present,future} tense verbs from the proximity analysis when the asset is the subject or object of the clause
GlobalProxSubObj	Global temporality by majority voting from the proximity analysis when the asset is the subject or object of the clause

- M is the size of the document collection.
- α and β are the symmetric smoothing hyper-parameters to avoid discarding one of the topics due to zero intra-document or intra-corpus topic occurrences. Both hyper-parameters are initialised randomly. Specifically, α and β are the topic-document and word-topic densities, respectively. Lower values of α and β reduce the variability of topic assignment to specific documents and words.
- $P(\theta_j; \alpha)$ and $P(\phi_i; \beta)$ are the topic-documents and word-topics Dirichlet distributions, respectively.
- $P(Z_{j,t}|\theta_j)$ and $P(W_{j,t}|\phi_{Z_{j,t}})$ are the topic-documents and word-topics multinomial distributions, respectively.

During LDA model training, by modifying α and β, (i) topics are assigned randomly to the words in each document, then, (ii) the algorithm iterates across the word-topic pairs in different documents generating new assignments and accepting them if they decrease the number of topics per word and document.

The algorithm converges when it finds a solution that minimises the number of intra-document topics and topics per word. Alternatively, an iteration limit can be set. Ultimately, the resulting assignment of words to topics can be used to define a criterion for detecting topics in new text. To this end, when a new sentence is presented to the algorithm, the step (ii) is repeated by also taking into account the words in the new sentence. The score of a sentence for a given topic is the number of words of that topic in the sentence divided by the length of the sentence in words. In principle, the sentence is considered to belong to the topic with the largest score. Note that in this estimation,

Table 8 Example of news entry after applying dependency and proximity analyses

Dependency analysis	Proximity analysis
TICKER (STOCK:TICKER_ABR) is proving to be a stable but undervalued company in the market today. In fact, **TICKER_ABR** stock **is** worth at least 55% more than TICKER_ABR stock price today using an analysis of TICKER_ABR stock dividend yield, TICKER_ABR stock own FIN_ABR ratio history, and a comparison with TICKER_ABR stock peers. TICKER reported "boring" earnings, according to Barron's magazine for Q2 on July 24. But Barron's magazine for Q2 says "boring is good" in this market. They point out that the communications business is not a bad place to be in a pandemic. For example, **TICKER said** that TICKER earnings on a non-FIN_ABR adjusted basis was $1.18 per share. This was only 3% lower than a year ago. On an adjusted FIN_ABR (earnings before interest, taxes, depreciation, and amortization) basis, taxes, depreciation, and amortization) cash flow was down 5%. However, **TICKER generated** higher cash flow. For example, first-half 2020 cash flow from operations of $23.6 billion, an increase of $7.7 billion from first-half of 2019. This represents a huge increase of over 206%. Moreover, this free cash flow (TICKER_ABR) in the first half was $13.7 billion, an increase of 74.1 percent year over year. But consider this. TICKER trades for a paltry 12.3 times this year's expected earnings and just 12 times next year	**TICKER (STOCK:TICKER_ABR) is proving to be** a stable but undervalued company in the market today. In fact, **TICKER_ABR** stock **is** worth at least 55% more than TICKER_ABR stock price today using an analysis of TICKER_ABR stock dividend yield, TICKER_ABR stock own FIN_ABR ratio history, and a comparison with TICKER_ABR stock peers. **TICKER reported** "boring" earnings, according to Barron's magazine for Q2 on July 24. But Barron's magazine for Q2 says "boring is good" in this market. They point out that the communications business is not a bad place to be in a pandemic. For example, **TICKER said** that TICKER earnings on a non-FIN_ABR adjusted basis was $1.18 per share. This was only 3% lower than a year ago. On an adjusted FIN_ABR (earnings before interest, taxes, depreciation, and amortization) basis, taxes, depreciation, and amortization) cash flow was down 5%. However, **TICKER generated** higher cash flow. For example, first-half 2020 cash flow from operations of $23.6 billion, an increase of $7.7 billion from first-half of 2019. This represents a huge increase of over 206%. Moreover, this free cash flow (**TICKER_ABR**) in the first half **was** $13.7 billion, an increase of 74.1 percent year over year. But consider this. TICKER trades for a paltry 12.3 times this year's expected earnings and just 12 times next year

the algorithm is started using the distribution with the best hyper-parameters α and β when the training algorithm terminates.

The capability of the system to differentiate between relevant and non-relevant information in the resulting topics is related to two combined effects of data conditioning in previous stages. First, TextTiling groups text by the different expression patterns that the authors of financial news tend to use in relevant and non-relevant text. Second, co-reference resolution and tag processing create a higher density of certain tags in relevant text. For this reason, as a practical contribution, we defined a topic score ρ that represents the density of significant tags STOCK, TICKER, CURRENCY and FIN_ABR in financial news content, which is computed as the percentage of significant tags in a topic divided by the total number of tags in the whole data set. The topic with the highest ρ value is considered relevant. Furthermore, to improve the precision of the LDA algorithm in detecting relevant text (that is, its ability to avoid false positives) we introduced another practical contribution: an LDA score threshold for accepting a sentence as relevant. It is computed as the minimum value of the configurable parameter δ and the mean value of the topic scores of the relevant sentences in the same segment.

Table 6 shows an example of relevant information detection, together with some sentences on the relevant topic from the same segment and the corresponding LDA classification scores. The mean value of the scores is 0.878.[5] Thus, assuming that $\delta = 0.8$, even though the third sentence belongs to the relevant topic, the system would consider it irrelevant.

3.5 Temporal analysis

Table 7 shows the set of temporal features used to train the ML temporal analysis model. The focal point of this analysis is the use of verbs when referring to stock markets, assets and currencies, but, unlike previous works, we consider them to be part of the semantic context.

Within each relevant sentence, we perform a dependency analysis to link verbs to stock markets, assets and currencies and a proximity analysis based on the proximity between the verbs and the key elements identified. The system measures the distance of a term to the nearest verb as the number of intermediate words in both directions. For both analyses (dependency and proximity), we consider whether assets are the subjects or objects of their clauses. For each feature, we estimate the verb tense by majority voting among past, present and future tenses (in case of a tie, the future tense prevails).

Algorithms 1 and 2 describe the generation of the temporal features of a segment based on the dependency (Algorithm 1) and the proximity (Algorithm 2) analysis of the corresponding sentences. Both algorithms have linear time complexity $O(n \cdot d)$ owing to the d independent loops with a limited number of n elements (subject and object sentences in our case). In the particular scenario in which the number of subject sentences equals the number of object sentences, for each loop and each tense, time complexity is $O(2 \cdot 3 \cdot n)$ for 2 analyses (subject and object) and 3 tenses (past, present, future). Table 8 shows an example of the outcome. Note the financial terms and associated verbs. For this segment, for example, features FutDepSubObj and FutProxSubObj in Table 7 are set to 1.

```
function DEPENDENCY_ANALYSIS(text)
    {Pst,Prs,Fut}DepSub = {Pst:0,Prs:0,Fut:0}
    {Pst,Prs,Fut}DepSubObj = {Pst:0,Prs:0,Fut:0}
    DepSub_text = parsing_subject(text)
    DepSubObj_text = parsing_subject_object(text)
    for tense in [Pst,Prs,Fut] do
        {Pst,Prs,Fut}DepSub.add(tense,
        get_number_verb_tense(DepSub_text,tense))
    end for
    GlobalDepSub = majority({Pst,Prs,Fut}DepSub)
    for tense in [Pst,Prs,Fut] do
        {Pst,Prs,Fut}DepSubObj.add(tense,
        get_number_verb_tense(DepSubObj_text,tense))
    end for
    GlobalDepSubObj = majority({Pst,Prs,Fut}
     DepSubObj)
end function
```

Algorithm 1 Dependency analysis

In addition to the temporal features, we also consider textual and numerical features in the ML temporal analysis model. The textual features are char-grams, word-grams and word tokens (n-grams within word boundaries), whose parameter ranges are selected by combinatorial searching. The two numerical features are the number of numerical values (excluding percentages) and the number of percentages in the news content.

After empirical tests with diverse ML algorithms, we chose a Linear Support Vector Classifier (SVC) to estimate the temporality (past, present, future) of a segment (see

[5] Note that this mean value is computed with all relevant sentences in the segment, only some of them are contained in Table 6.

```
function PROXIMITY_ANALYSIS(text)
    {Pst,Prs,Fut}ProxSub = {Pst:0,Prs:0,Fut:0}
    {Pst,Prs,Fut}ProxSubObj = {Pst:0,Prs:0,Fut:0}
    ProxSub_text = parsing_subject(text)
    ProxSubObj_text = parsing_subject_object(text)
    for tense in [Pst,Prs,Fut] do
        {Pst,Prs,Fut}ProxSub.addNearest(tense,
        get_number_verb_tense(ProxSub_text,tense))
    end for
    GlobalProxSub = majority({Pst,Prs,Fut}ProxSub)
    for tense in [Pst,Prs,Fut] do
        {Pst,Prs,Fut}ProxSubObj.addNearest(tense,
        get_number_verb_tense(ProxSubObj_text,tense))
    end for
    GlobalProxSubObj = majority({Pst,Prs,Fut}
     ProxSubObj)
end function
```

Algorithm 2 Proximity analysis

our previous work [61]). Before training the SVC, we pre-processed the clauses in the financial news by converting text to lower case, and removing punctuation marks and non-Unicode characters such as accents and symbols.

Finally, Algorithm 3 shows the logical flow of the proposed solution.

4 Experimental results

In this section, we describe the experimental data set and the performance of our system for the detection of relevant financial events. Well-known state-of-the-art metrics were used for the evaluation: Alpha-reliability and accuracy [62], and the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric[6] [21, 63, 64]. Comparisons of the proposed solution with a rule-based baseline and a supervised extraction approach are also provided.

4.1 Experimental setting

The experiments were performed on a computer with the following specifications:

- Operating system: Ubuntu 18.04.2 LTS 64 bits
- Processor: IntelCore i9-9900K 3.60 GHz
- RAM: 32 GB DDR4
- Disk: 500 GB (7200 rpm SATA) + 256 GB SSD

Regarding the implementation, we took the following decisions:

```
W: 20
K: 10
similarity_method: block comparison
stopwords_list: nltk
smoothing_width: 2
smoothing_rounds: 1
```

Listing 1 Configuration parameter ranges of the TextTiling algorithm

- Segmentation (Section 3.1): as previously mentioned, we applied the TextTiling algorithm.[7] For this first stage, we used the default parameters of the classifier (see Listing 1). Stop-words were also removed.
- Co-reference detection (Section 3.2): we used the NN implementation of the NeuralCoref library[8], which is a pipelined extension of the spaCy library[9] to solve co-reference groups in blocks of text. Specifically, we used the pre-trained word embedding statistical model for English with its default features. Surprisingly, in our analysis, we noticed that NeuralCoref did not always detect the indispensable word *stock*. To circumvent this problem, following a trial and error process, we replaced *stock* with *index* before co-reference resolution.
- Tag processing (Section 3.3): we employ the Freeling library[10] to detect dependencies between terms. Capitalised proper names are detected with the spaCy[11] library. If the names are not in our lexica, their categories are selected from spaCy EntityRecognizer tool.[12] Depending on the category, the names are replaced by the following tags: MONEY, PERSON, NORP (nationalities, religions or political groups), ORG (organisations, companies), PRODUCT, EVENT, and WORK OF ART (titles of artworks). Moreover, all elements recognised as LOC, FAC (buildings) and GPE (countries, cities, states) are grouped under the tag LOC (locations). Numerical values and dates are detected using EntityRecognizer. Terms recognised as DATE and TIME are grouped under the tag DATE, while those recognised as PERCENT, CARDINAL and QUANTITY are grouped under the tag NUM.
- Relevant text detection with LDA topic modelling (Section 3.4): we employed the LdaMulticore module

[6] Available at https://github.com/pltrdy/rouge, October 2022.

[7] Available at https://www.nltk.org/_modules/nltk/tokenize/texttiling.html, October 2022.

[8] Available at https://github.com/huggingface/neuralcoref, October 2022.

[9] Available at https://spacy.io, October 2022.

[10] Available at http://nlp.lsi.upc.edu/freeling/node/1, October 2022.

[11] Available at https://spacy.io, October 2022.

[12] Available at https://spacy.io/api/top-level#spacy.explain and https://github.com/explosion/spaCy/blob/master/spacy/glossary.py, October 2022.

```
function RELEVANT_FORECAST_DETECTION(financial_news)
    segments = TextTiling(financial_news) %Multi-paragraph topic segmentation
    for s in segments do
        NeuralCoref(s) %Co-reference resolution
        %Financial terms detection, homogenisation and replacement
        replaceAll(s,stock_dict,“STOCK”)
        replaceAll(s,asset_dict,“TICKER”)
        keywords = [“company”,“enterprise”,“manufacturer”,“shareholder”]
        replaceAll(s,keywords,“TICKER”)
        replaceAll(s,assetsAbr_dict,“TICKER_ABR”)
        replaceAll(s,currency_dict,“CURRENCY”)
        replaceAll(s,financial_dict,“FIN_ABR”)
        NER(s)
    end for
    topic_modelling = LDA(segments) %Relevant text detection with LDA topic modelling
    relevant_topic=0
    highest_$\rho$ = 0
    for topic in topics do
        topic_$\rho$ = max(topic_modelling.filterTopic[topic].$\rho$)
        if topic_$\rho$ > highest_$\rho$ then
            relevant_topic = topic
            highsest_$\rho$ = topic_$\rho$
        end if
    end for
    relevant = topic_modelling.filterTopic[relevant_topic]
    for sentence in relevant do
        if sentence.score < $\rho$ then
            relevant.remove(sentence)
        end if
    end for
    %Temporal analysis
    for s in segments do
        to_lowercase(s)
        remove_punctuation_non-Unicode(s)
        dependency = dependency_analysis(s)
        proximity = proximity_analysis(s)
        textual = compute_chargrams_wordgrams_tokens(s)
        numerical = count_numerical_values_percentages(s)
    end for
    trainSVC(dependency,proximity,textual,numerical)
end function
```

Algorithm 3 Solution pipeline

from the gensim Python library.[13] Listing 2 shows the configuration parameters used to train the model. Through repeated trials, we tuned the algorithm to 50 training `passes`, `alpha` to `symmetric` and `beta` to `asymmetric`. Finally, we set $\delta = 0.8$.

- `Numtopics` is the number of latent topics to be extracted.
- `Passes` is the number of passes to be applied during training.
- `Random state` is a useful seed for reproducibility.
- `Alpha` represents an a-priori belief about document-topic distribution, that is, prior to selection strategies. Its feasible values are: (*i*) `scalar` for symmetric document-topic distribution, (*ii*) `symmetric` to use a fixed

[13] Available at https://pypi.org/project/gensim, October 2022.

Listing 2 Configuration parameter ranges for the LDA model

```
numtopics: 2
passes: (25,50,75,100)
randomstate: 1
alphabeta: (0.01,0.31,0.61,0.91,symmetric,asymmetric)
```

symmetric distribution of $1.0/numtopics$, and (*iii*) `asymmetric` to use a fixed normalised asymmetric distribution of $1.0/(topicindex + sqrt(numtopics))$.

  - `Beta` is an a-priori belief on topic-word distribution. It has the same feasible values as `alpha`.

- Temporal analysis (Section 3.5): Freeling is used to tag assets as subjects or objects of their clauses to obtain the corresponding temporal features. We used the SVC implementation from the Scikit-Learn Python library.[14] Regarding the parameterisation of the char-grams, word-grams and word tokens textual features, we applied a GridSearchCV[15] combinatorial search from Scikit-Learn within the ranges in our prior related work [61]. The final choices were `maxdf` = 0.30, `mindf` = 0, `ngramrange` = (2,4) and `maxfeatures` = 10000.
  - `Mindf` and `maxdf` are used to ignore terms with a lower (cut-off) and higher (corpus-specific stop words) document frequency than the given threshold, respectively.
  - `Ngramrange` indicates the lower and upper boundary for the extraction of word *n*-grams.
  - `Maxfeatures` represents the number of features considered for the best split.

  To select the best features for the temporal analysis across the whole set (temporal, textual and numerical features), we used SelectPercentile[16] from Scikit-Learn with the χ^2 score function and 80th percentile threshold. The hyper-parameters of the SVC were tuned using GridSearchCV with 10-fold cross validation within the ranges in Listing 3. The optimal hyper-parameter values used were `C` = 0.001, `classweight` = balanced, `loss` = `squared_hinge`, `maxiter` = 1500, `multiclass` = ovr, `penalty` = l2, `tol` = 10^{-9}. Finally, the SVC was evaluated by 10-fold cross validation using 600 sentences from financial news. This auxiliary data set is similar in size to other sets used in the literature [26–31] and did not belong to the experimental data set described in Section 4.1 that was used to detect relevant information, and was independently annotated. The SVC attained 80.21% precision and 80.40% recall for the auxiliary set.

  - `C` is the regularisation parameter.
  - `Classweight` is used to set the parameter `C` of the classes. If not given, all classes are assumed to have weight one. The `balanced` mode automatically adjusts weights in a manner that is inversely proportional to class frequencies in the input data.
  - `Loss` represents the loss function.
  - `Maxiter` is the hard limit on iterations, or -1 for no limit.
  - `Multiclass` represents the one-vs-one scheme.
  - `Penalty` represents the penalty for the model.
  - `Tol` represents the tolerance for the stopping criterion.

4.2 Experimental data set

Our experimental data set was composed of 2,158 news pieces (average length of 27.98 sentences and 537.24 words). As previously mentioned, the pieces were automatically extracted with a script from popular and prestigious financial websites between 1st October 2018 and 1st October 2020. We filtered the news pieces to keep those that mentioned at least one of the stocks in our financial lexica.[3] For text processing purposes, double spaces, line breaks and tabs were replaced by a single space. Finally, we removed, URLs, images and graphics and kept the date and author information.

Each entry in the resulting data set is composed of an identifier, a title, content, author information, source and date of publication. The entries are comparable in size to those described in previous KE [26–31] and LDA works [65].

A brief descriptive analysis of the data set is given in Table 9.

The texts were annotated by five NLP scientists from the atlanTTic Research Centre for Telecommunication Technologies at the University of Vigo. Manual annotations included relevant texts, asset identifiers and prediction/forecast texts. A number of guidelines were agreed on to enhance consistency (*e.g.*, bold font for relevant

[14] Available at https://scikit-learn.org/stable/supervised_learning.html#supervised-learning, October 2022.

[15] Available at https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html , October 2022.

[16] Available at https://scikit-learn.org/stable/modules/feature_selection.html , October 2022.

Listing 3 Configuration parameter ranges for SVC

```
C: (0.0001, 0.0005, 0.001, 0.01)
classweight: [None,balanced]
loss: squared_hinge
maxiter: [500, 1000, 1500]
multiclass: ovr
penalty: l2
tol: (0.000000001, 0.00000001, 0.0000001, 0.000001)
```

text, italics for asset identifiers and underlining for prediction/forecast text). Table 10 shows an example of an annotated news item from the experimental data set. We used the annotated asset identifiers to improve the content of our financial lexica[3] which increased by 3.95%.

4.3 Inter-agreement evaluation

We evaluated inter-annotator agreement using two well-known state-of-the-art metrics: Alpha-reliability and accuracy.

Table 9 Descriptive analysis of the experimental data set

Source	Authors	News	Avg. chars	Avg. words	
	Benzinga[a]	152	400	3325.21	536.295
	The Motley Fool[b]	75	298	3537.58	576.69
	Markets Insider[c]	13	223	5270.50	758.61
	TipRanks[d]	4	210	1505.61	235.84
	Business Standard[e]	22	198	2615.02	430.00
	Investorplace[f]	42	158	4967.15	827.32
	Iwatch Markets[g]	2	154	555.51	95.24
	Investing Daily[h]	2	150	6319.58	1011.69
	Investopedia[i]	17	140	4462.16	703.70
	Gurufocus[j]	20	49	4815.94	783.65
	CNN Business[k]	32	49	4507.49	745.49
	Market Watch[l]	28	47	4352.68	721.57
	Seeking Alpha[m]	13	33	783.67	124.45
	CNBC[n]	4	31	1962.16	338.87
	IOL[o]	6	18	1046.00	169.22
Total	15	431[p]	2158	3335.08	537.24

[a] Available at https://www.benzinga.com, October 2022

[b] Available at https://www.fool.com, October 2022

[c] Available at https://markets.businessinsider.com, October 2022

[d] Available at https://www.tipranks.com, October 2022

[e] Available at https://www.business-standard.com, October 2022

[f] Available at https://investorplace.com, October 2022

[g] Available at https://iwatchmarkets.com, October 2022

[h] Available at https://www.investingdaily.com, October 2022

[i] Available at https://www.investopedia.com, October 2022

[j] Available at https://www.gurufocus.com, October 2022

[k] Available at https://edition.cnn.com/markets, October 2022

[l] Available at https://www.marketwatch.com, October 2022

[m] Available at https://seekingalpha.com, October 2022

[n] Available at https://www.cnbc.com, October 2022

[o] Available at https://www.iol.co.za, October 2022

[p] Without duplicates

Table 10 Example of annotated news entry

Verizon Communications (NYSE:VZ) is proving to be a stable but undervalued company in the market today. **In fact, *VZ* stock is worth at least 55% more than its price today using an analysis of its dividend yield, its own P/E ratio history, and a comparison with its peers.** The company reported "boring" earnings, according to Barron's magazine for Q2 on July 24. But the magazine says "boring is good" in this market. They point out that the communications business is not a bad place to be in a pandemic. For example, Verizon said that its earnings on a non-GAAP adjusted basis was $1.18 per share. This was only 3% lower than a year ago. On an adjusted EBITDA (earnings before interest, taxes, depreciation, and amortization) basis, its cash flow was down 5%. However, *Verizon* generated higher cash flow. For example, first-half 2020 cash flow from operations of $23.6 billion, an increase of $7.7 billion from first-half of 2019. This represents a huge increase of over 206%. Moreover, its free cash flow (FCF) in the first half was $13.7 billion, an increase of 74.1 percent year over year. But consider this. ***Verizon*** **trades for a paltry 12.3 times this year's expected earnings and just 12 times next year[...]**

Table 11 shows the coincidence matrix of relevance across all annotators. The two components in the diagonal show the number of news sentences on which all the annotators agreed, while the other two components show the cases on which at least one annotator disagreed. Tables 12 and 13 show the Alpha-reliability and accuracy coefficients by pairs of annotators. The mean values were 0.552 and 0.861, respectively. Previous works have considered an Alpha-reliability value above 0.41 to be acceptable [66–69]. Inter-agreement accuracy was very high, at over 80%.

4.4 Discussion of the results

Before applying our system for the automatic detection of relevant financial events, we first defined a simple rule-based system as a baseline. This system sets a relevance score by counting tickers, numbers and percentages and detecting future tenses using Freeling. As previously mentioned, relevant financial text has characteristic context data and expression patterns, and our goal was to determine whether a sophisticated technique such as LDA would perform better than trivial rules.

The rule-based baseline approach is as follows. First, the text is segmented into sentences, and all references to stock markets, assets, asset abbreviations and currencies are replaced by the tags STOCK, TICKER, TICKER_ABR and CURRENCY, respectively. As indicated in Section 3.3, financial terms and abbreviations are also replaced by the tag FIN_ABR. Freeling is applied to detect percentages and numerical values, which are replaced by the tag NUM. Sentences containing a future tense as detected by Freeling are considered to refer to the future. In brief, a sentence is considered relevant if it contains at least one financial tag (STOCK, TICKER, TICKER_ABR, CURRENCY or FIN_ABR) and at least one NUM tag, and predictive if the main verb is in future tense.

Table 11 Coincidence matrix for relevant text annotation

	Relevant	Context
Relevant	2752.5	1561.5
Context	1561.5	16584.5

Drawing from related work on more powerful supervised extraction strategies [43–45, 49], we also applied a SVC model[17] as a second comparison reference. The model was trained using manual annotations on relevant text, including predictions. Textual features were generated and hyper-parameter settings optimised as described in Section 4.1.

Next, we evaluated our system by checking it against the annotated segments. To do this, we employed ROUGE, a widely used set of metrics for evaluating automatic text extraction performance based on overlapping n-grams. We used ROUGE-L, which measures the longest common sub-sequence between the system output and the annotated news. This ROUGE variant has been applied as a string matching algorithm to compute the similarity between two texts [70].

Tables 14 and 15 show the results obtained for the baseline systems and the proposed system. Even though the problem case is entirely novel (see Section 1.5), the results show that the application of sophisticated NLP techniques and KE algorithms such as those used in our solution results in improved extraction of relevance and temporality from financial news content.

Table 14 shows the results for the detection of relevant information by the baseline systems and our system for the tags identified by the five annotators. Average values are also provided. In our tests, text was considered relevant when its score for a given topic doubled the score of the other topic.[18] The remaining text was considered to be less relevant or contextual information. The average ROUGE-L value across all annotators was 0.662, more than doubling the performance of the rule-based baseline approach. The manually intensive supervised extraction alternative was comparable to our unsupervised approach (in fact the former was often worse, depending on the annotator). This performance can be considered satisfactory in line with other works from the literature [63, 64, 71, 72]. Table 15 shows the results for the detection of relevant predictions/forecasts after the last SVC classifier stage. The average ROUGE-L value in this case, 0.982, was excellent,

[17] Note that a supervised approach was also used to estimate the temporality (past, present, future) of text in our proposal.

[18] Note that there are only two topics under analysis in this work.

Table 12 Inter-agreement Alpha-reliability of relevant text annotation by pairs of annotators (An.)

	An. 1	An. 2	An. 3	An. 4	An. 5
An. 1	–	0.569	0.550	0.605	0.629
An. 2	0.569	–	0.594	0.525	0.520
An. 3	0.550	0.594	–	0.436	0.506
An. 4	0.605	0.525	0.436	–	0.585
An. 5	0.629	0.520	0.506	0.585	–

Table 13 Inter-agreement accuracy of relevant text annotation by pairs of annotators (An.)

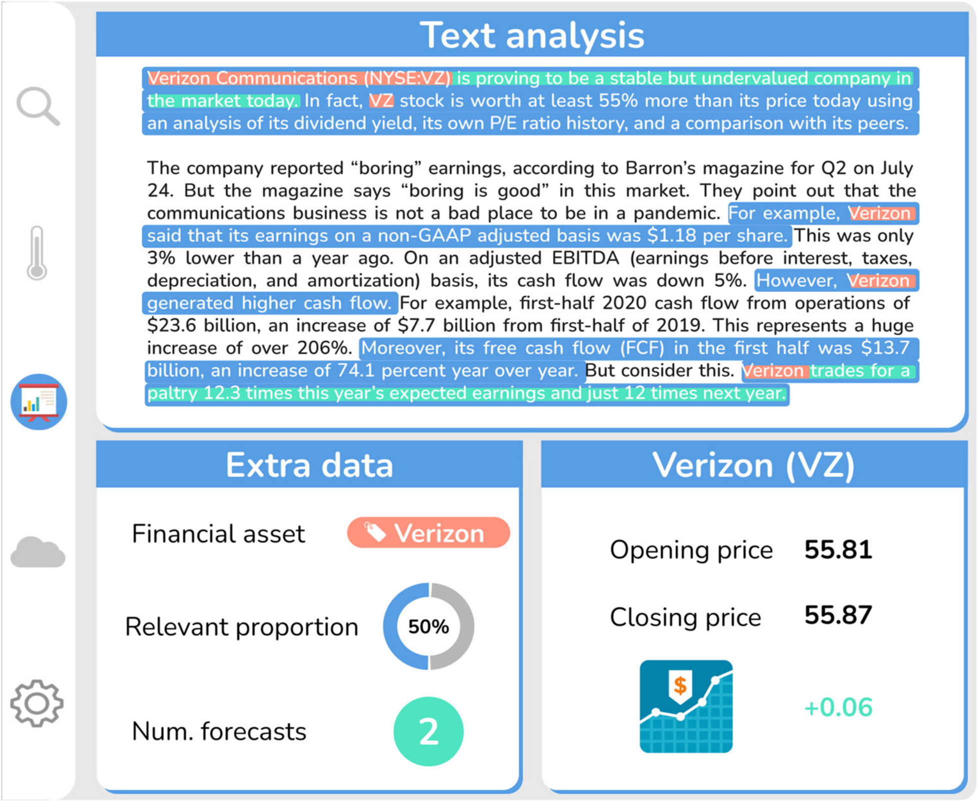
	An. 1	An. 2	An. 3	An. 4	An. 5
An. 1	–	0.862	0.844	0.891	0.895
An. 2	0.862	–	0.853	0.859	0.855
An. 3	0.844	0.853	–	0.818	0.838
An. 4	0.891	0.859	0.818	–	0.894
An. 5	0.895	0.855	0.838	0.894	–

Table 14 ROUGE-L measures for the detection of relevant text, by annotator (An.) and average

	An. 1	An. 2	An. 3	An. 4	An. 5	Avg.
Rule-based baseline	0.257	0.390	0.333	0.390	0.246	0.323
Supervised system	0.655	0.582	0.615	0.643	0.574	0.614
Proposed system	0.727	0.681	0.720	0.547	0.633	0.662

Table 15 ROUGE-L measures the detection of relevant text with predictions/forecasts, by annotator (An.) and average

	An. 1	An. 2	An. 3	An. 4	An. 5	Avg.
Rule-based baseline	0.748	0.716	0.661	0.715	0.723	0.713
Supervised system	0.984	0.905	0.906	0.938	0.945	0.936
Proposed system	0.991	0.970	0.975	0.982	0.990	0.982



Verizon Communications (NYSE:VZ) is proving to be a stable but undervalued company in the market today. In fact, VZ stock is worth at least 55% more than its price today using an analysis of its dividend yield, its own P/E ratio history, and a comparison with its peers.

The company reported "boring" earnings, according to Barron's magazine for Q2 on July 24. But the magazine says "boring is good" in this market. They point out that the communications business is not a bad place to be in a pandemic. For example, Verizon said that its earnings on a non-GAAP adjusted basis was $1.18 per share. This was only 3% lower than a year ago. On an adjusted EBITDA (earnings before interest, taxes, depreciation, and amortization) basis, its cash flow was down 5%. However, Verizon generated higher cash flow. For example, first-half 2020 cash flow from operations of $23.6 billion, an increase of $7.7 billion from first-half of 2019. This represents a huge increase of over 206%. Moreover, its free cash flow (FCF) in the first half was $13.7 billion, an increase of 74.1 percent year over year. But consider this. Verizon trades for a paltry 12.3 times this year's expected earnings and just 12 times next year.

Financial asset Verizon

Relevant proportion 50%

Num. forecasts 2

Opening price 55.81

Closing price 55.87

+0.06

Fig. 2 Example of financial event detection

with a significant improvement over the rule-based baseline reference of 0.713.

Counterintuitively, the performance of the rule-based baseline in Table 14 is worse because this approach gives relevance to text that contains quantitative data even if it corresponds to merely contextual information, such as past states of assets and stock markets. It underperformed our proposed solution by an average of 51%. The level of agreement between our system and the annotators for the detection of predictions/forecasts (Table 15) was near perfect, with ROUGE values of more than 0.970 for all annotators. As expected, predictions and forecasts within relevant text are easier to detect than relevant text itself, explaining the lower ROUGE values in this second case, where the highest coefficient observed was 0.727 (for annotator 1).

4.5 Application use case

Figure 2 shows a news piece highlighted by our system. Relevant sentences are highlighted in blue, asset identifiers in pink and predictions/forecasts in green. Note the differences with the manual highlighting in Table 10. For example, human annotators might mark the sentence "In fact, VZ stock is worth at least 55% more than its price today" as a prediction. With our system, however, both forecasts and relevant, informative sentences are marked.

The dashboard at the bottom of Fig. 2 summarises the results, showing the proportion of relevant segments and number of predictions and forecasts. The updated value of the financial asset, taken from Yahoo Finance[19], is shown on the right.

5 Conclusions

Many valuable online financial news sources, such as economy journals and web pages (Motley Fool, InvestorDaily, etc.), contain opinions from experts describing relevant market events within sociological, political and/or cultural contexts.

The system proposed in this paper is designed to extract this relevant information, and in particular forecasts and predictions. To do this, it employs NLP techniques. It segments the text and applies LDA analysis to filter out less relevant sentences, and then applies discursive temporality analysis to identify predictions and forecasts within the remaining relevant text. The result is a summary of relevant, easy-to-read information. We are not aware of any other KE systems that have applied a similar approach to resolve this problem.

To our knowledge and considering related work, our proposal is the first to jointly consider relevance and

[19] Available at https://finance.yahoo.com, October 2022.

temporality at the discursive level. It contributes to transferring human associative discourse capabilities to expert systems by combining (*i*) multi-paragraph topic segmentation and co-reference resolution to separate author expression patterns, (*ii*) detection of relevant text through topic modelling with LDA, and (*iii*) identification of forecasts and predictions within relevant text using discursive temporality analysis and ML.

We have created an experimental data set composed of 2,158 financial news items to evaluate our proposal. We have validated its annotation capacity by performing an inter-agreement analysis using Alpha-reliability and accuracy measures and evaluated its performance using the state-of-the-art ROUGE metric. The system attained ROUGE-L values of 0.662 and 0.982 for the detection of relevant data and predictions/forecasts, respectively. We also compared the performance of our system with a rule-based baseline system and a fully supervised system (which also performs supervised extraction of relevant text) to evaluate its competitiveness. It outperformed the rule-based system and was comparable to the fully supervised system, which unlike our solution requires manual annotation.

In future work, we plan to extend our research to Spanish and other languages to cover a broader community of investors. We will also evaluate the system in composite (multi-disciplinary) research domains.

Author Contributions **Silvia García-Méndez**: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - original draft. **Francisco de Arriba-Pérez**: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - original draft. **Ana Barros-Vila**: Software, Validation, Investigation, Resources, Data Curation, Writing - review & editing. **Francisco J. González-Castaño**: Conceptualisation, Methodology, Data Curation, Writing - review & editing, Supervision. **Enrique Costa-Montenegro**: Methodology, Data Curation, Writing - review & editing.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was partially supported by Xunta de Galicia grants ED481B-2021-118 and ED481B-2022-093, Spain. University of Vigo/CISUG covered the open access fee.

Data Availability The data sets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Code Availability The code used in this work is not publicly available.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for Publication Not applicable.

References

1. Manogaran G, Varatharajan R, Lopez D et al (2018) A new architecture of internet of things and big data ecosystem for secured smart healthcare monitoring and alerting system. Futur Gener Comput Syst 82:375–387. https://doi.org/10.1016/j.future.2017.10.045
2. Delić V, Perić Z, Sečujski M et al (2019) Speech technology progress based on new machine learning paradigm. Comput Intell Neurosci 2019:1–19. https://doi.org/10.1155/2019/4368036
3. Ma X, Fei Q, Qin H et al (2020) A new efficient decision making algorithm based on interval-valued fuzzy soft set. Appl Intell 51(6):3226–3240. https://doi.org/10.1007/s10489-020-01915-w
4. Zuo Y, Wu Y, Min G et al (2020) An intelligent anomaly detection scheme for micro-services architectures with temporal and spatial data analysis. IEEE Trans Cogn Commun Netw 6(2):548–561. https://doi.org/10.1109/TCCN.2020.2966615
5. Guetterman TC, Chang T, DeJonckheere M et al (2018) Augmenting qualitative text analysis with natural language processing: methodological study. J Med Int Res 20(6):e231. https://doi.org/10.2196/jmir.9702
6. Zhang F, Fleyeh H, Wang X, et al. (2019) Construction site accident analysis using text mining and natural language processing techniques. Autom Constr 99:238–248. https://doi.org/10.1016/j.autcon.2018.12.016
7. Balyan R, McCarthy KS, McNamara DS (2020) Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. Int J Artif Intell Educ 30(3):337–370. https://doi.org/10.1007/s40593-020-00201-7
8. Lu X, Deng Y, Sun T et al (2022) MKPM: multi keyword-pair matching for natural language sentences. Appl Intell 52(2):1878–1892. https://doi.org/10.1007/s10489-021-02306-5
9. Kumar S, Kumar MA, Soman K (2019) Deep learning based part-of-speech tagging for Malayalam twitter data (special issue: deep learning techniques for natural language processing). J Intell Syst 28(3):423–435. https://doi.org/10.1515/jisys-2017-0520
10. K. V, Gupta D (2018) Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: comparisons, analysis and challenges. Inf Process Manag 54(3):408–432. https://doi.org/10.1016/j.ipm.2018.01.008
11. Xing FZ, Cambria E, Welsch RE (2018) Natural language based financial forecasting: a survey. Artif Intell Rev 50(1):49–73. https://doi.org/10.1007/s10462-017-9588-9
12. Lytos A, Lagkas T, Sarigiannidis P et al (2019) The evolution of argumentation mining: from models to social media and emerging tools. Inf Process Manag 56(6):102,055. https://doi.org/10.1016/j.ipm.2019.102055
13. Kelly S, Ahmad K (2018) Estimating the impact of domain-specific news sentiment on financial assets. Knowl-Based Syst 150:116–126. https://doi.org/10.1016/j.knosys.2018.03.004

14. Atkins A, Niranjan M, Gerding E (2018) Financial news predicts stock market volatility better than close price. J Financ Data Sci 4(2):120–137. https://doi.org/10.1016/j.jfds.2018.02.002
15. Isa K, Rahman Ahmad A, Md Yusoff R et al (2018) NEWS analysis towards youth financial competency management. Int J Eng Technol 7(2.29):1151. https://doi.org/10.14419/ijet.v7i2.29.15146
16. Zhang H, Boons F, Batista-Navarro R (2019) Whose story is it anyway? Automatic extraction of accounts from news articles. Inf Process Manag 56(5):1837–1848. https://doi.org/10.1016/j.ipm.2019.02.012
17. Cepoi CO (2020) Asymmetric dependence between stock market returns and news during COVID-19 financial turmoil. Financ Res Lett 36:101,658. https://doi.org/10.1016/j.frl.2020.101658
18. Swathi T, Kasiviswanath N, Rao AA (2022) An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. Appl Intell :1–14
19. Loughran T, McDonald B (2016) Textual analysis in accounting and finance: a survey. J Account Res 54(4):1187–1230. https://doi.org/10.1111/1475-679X.12123
20. Lutz B, Pröllochs N, Neumann D (2020) Predicting sentence-level polarity labels of financial news using abnormal stock returns. Exp Syst Appl 148:113,223. https://doi.org/10.1016/j.eswa.2020.113223
21. Mohamed M, Oussalah M (2019) SRL-ESA-TextSum: a text summarization approach based on semantic role labeling and explicit semantic analysis. Inf Process Manag 56(4):1356–1372. https://doi.org/10.1016/j.ipm.2019.04.003
22. Evers-Vermeul J, Hoek J, Scholman MC (2017) On temporality in discourse annotation: Theoretical and practical considerations. Dialogue Discourse 8(2):1–20. https://doi.org/10.5087/dad.2017.201
23. Jang Y, Park CH, Seo YS (2019) Fake news analysis modeling using quote retweet. Electronics 8(12):1377. https://doi.org/10.3390/electronics8121377
24. Chau JY, Reyes-Marcelino G, Burnett AC et al (2019) Hyping health effects: a news analysis of the ‘new smoking’ and the role of sitting. Br J Sports Med 53(16):1039–1040. https://doi.org/10.1136/bjsports-2018-099432
25. Phi GT (2020) Framing overtourism: a critical news media analysis. Curr Issues Tour 23(17):2093–2097. https://doi.org/10.1080/13683500.2019.1618249
26. Li Y, Pan Q, Wang S et al (2018) A Generative model for category text generation. Inf Sci 450:301–315. https://doi.org/10.1016/j.ins.2018.03.050
27. Long W, Song L, Tian Y (2019) A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity. Exp Syst Appl 118:411–424. https://doi.org/10.1016/j.eswa.2018.10.008
28. Al-Smadi M, Al-Ayyoub M, Jararweh Y et al (2019) Enhancing aspect-based sentiment analysis of Arabic Hotels’ reviews using morphological, syntactic and semantic features. Inf Process Manag 56(2):308–319. https://doi.org/10.1016/j.ipm.2018.01.006
29. Zhang X, Ghorbani AA (2020) An overview of online fake news: characterization, detection, and discussion. Inf Process Manag 57(2):102,025. https://doi.org/10.1016/j.ipm.2019.03.004
30. de Oliveira Carosia AE, Coelho GP, da Silva AEA (2021) Investment strategies applied to the Brazilian stock market: a methodology based on sentiment analysis with deep learning. Exp Syst Appl 184:115,470. https://doi.org/10.1016/j.eswa.2021.115470
31. Xie M, Ye Z, Pan G et al (2021) Incomplete multi-view subspace clustering with adaptive instance-sample mapping and deep feature fusion. Appl Intell 51(8):5584–5597. https://doi.org/10.1007/s10489-020-02138-9
32. Nti IK, Adekoya AF, Weyori BA (2020) A systematic review of fundamental and technical analysis of stock market predictions. Artif Intell Rev 53(4):3007–3057. https://doi.org/10.1007/s10462-019-09754-z
33. Carta S, Corriga A, Ferreira A et al (2021) A multilayer and multi-ensemble stock trader using deep learning and deep reinforcement learning. Appl Intell 51(2):889–905. https://doi.org/10.1007/s10489-020-01839-5
34. Khan W, Ghazanfar MA, Azam MA et al (2022) Stock market prediction using machine learning classifiers and social media, news. J Ambient Intell Humanized Comput 13(7):3433–3456. https://doi.org/10.1007/s12652-020-01839-w
35. Rustam F, Reshi AA, Mehmood A et al (2020) COVID-19 future forecasting using supervised machine learning models. IEEE Access 8:101,489–101,499. https://doi.org/10.1109/ACCESS.2020.2997311
36. Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF (2020) A review of unsupervised feature selection methods. Artif Intell Rev 53(2):907–948. https://doi.org/10.1007/s10462-019-09682-y
37. García-Méndez S, Fernández-Gavilanes M, Juncal-Martínez J et al (2020) Identifying banking transaction descriptions via support vector machine short-text classification based on a specialized labelled corpus. IEEE Access 8:61,642–61,655. https://doi.org/10.1109/ACCESS.2020.2983584
38. De Arriba-Pérez F, García-Méndez S, Regueiro-Janeiro JA et al (2020) Detection of financial opportunities in micro-blogging data with a stacked classification system. IEEE Access 8:215,679–215,690. https://doi.org/10.1109/ACCESS.2020.3041084
39. Beliga S, Meštrović A, Martinčić-Ipšić S (2015) An overview of graph-based keyword extraction methods and approaches. J Inf Organ Sci 39(1):1–20
40. Kaiser K, Miksch S (2005) Information extraction. A survey. Tech. rep., Institute of Software Technology & Interactive Systems, Vienna University of Technology
41. Li C, Guo J, Lu Y et al (2018) LDA Meets Word2Vec. In: Proceedings of the The Web Conference. ACM Press, pp 1699–1706, https://doi.org/10.1145/3184558.3191629
42. Azhari M, Kumar YJ (2017) Improving text summarization using neuro-fuzzy approach. J Inf Telecommun 1(4):1–14. https://doi.org/10.1080/24751839.2017.1364040
43. Gottipati S, Shankararaman V, Lin JR (2018) Text analytics approach to extract course improvement suggestions from students’ feedback. Res Pract Technol Enhanc Learn 13(1):6. https://doi.org/10.1186/s41039-018-0073-0
44. López-Úbeda P, Díaz-Galiano MC, Ureña-López LA et al (2021) Pre-trained language models to extract information from radiological reports. In: CEUR Workshop Proceedings, vol 2936. CEUR
45. Vermeer SA, Araujo T, Bernritter SF et al (2019) Seeing the wood for the trees: how machine learning can help firms in identifying relevant electronic word-of-mouth in social media. Int J Res Mark 36(3):492–508. https://doi.org/10.1016/j.ijresmar.2019.01.010
46. Jacobs G, Lefever E, Hoste V (2018) Economic event detection in company-specific news text. In: Proceedings of the first workshop on economics and natural language processing. association for computational linguistics, pp 1–10, https://doi.org/10.18653/v1/W18-3101
47. Oncharoen P, Vateekul P (2018) Deep learning for stock market prediction using event embedding and technical indicators. In: Proceedings of the international conference on advanced

informatics: concept theory and applications. IEEE, pp 19–24, https://doi.org/10.1109/ICAICTA.2018.8541310

48. Carta S, Consoli S, Piras L et al (2021) Event detection in finance using hierarchical clustering algorithms on news and tweets. PeerJ Comput Sci 7:e438. https://doi.org/10.7717/peerj-cs.438
49. Harb A, Plantié M, Dray G et al (2008) Web opinion mining. In: Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology. ACM Press, p 211, https://doi.org/10.1145/1456223.1456269
50. Shilpa B, Shambhavi B (2021) Combined deep learning classifiers for stock market prediction: integrating stock price and news sentiments. Kybernetes pp 1–26
51. Genç S, Akay D, Boran FE et al (2020) Linguistic summarization of fuzzy social and economic networks: an application on the international trade network. Soft Comput 24(2):1511–1527. https://doi.org/10.1007/s00500-019-03982-9
52. Abu El-Qumsan AY, El-Halees AM (2018) Template based medical reports summarization. Int J Comput Appl 179(17):47–55. https://doi.org/10.5120/ijca2018916301
53. Meena YK, Gopalani D (2020) Statistical features for extractive automatic text summarization. In: Natural language processing: concepts, methodologies, tools, and applications. IGI Global, pp 619–637, https://doi.org/10.4018/978-1-7998-0951-7.ch030
54. Gupta S, Gupta SK (2019) Abstractive summarization: an overview of the state of the art. Exp Syst Appl 121:49–65. https://doi.org/10.1016/j.eswa.2018.12.011
55. Alhoshan M, Altwaijry N (2020) AUSS: an Arabic query-based update-summarization system. J King Saud Univ Comput Inf Sci 1:1319–1578. https://doi.org/10.1016/j.jksuci.2020.11.027
56. Barros C, Lloret E, Saquete E et al (2019) NATSUM: narrative abstractive summarization through cross-document timeline generation. Inf Process Manag 56(5):1775–1793. https://doi.org/10.1016/j.ipm.2019.02.010
57. He X, Wang J, Zhang Q et al (2020) Improvement of text segmentation texttiling algorithm. J Phys Conf Ser 1453:12,008–12,015. https://doi.org/10.1088/1742-6596/1453/1/012008
58. Clark K, Manning CD (2016) Improving coreference resolution by learning entity-level distributed representations. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for Computational Linguistics, pp 643–653, https://doi.org/10.18653/v1/P16-1061
59. Jelodar H, Wang Y, Yuan C et al (2019) Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 78(11):15,169–15,211. https://doi.org/10.1007/s11042-018-6894-4
60. Gupta A, Katarya R (2021) PAN-LDA: a latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning. Comput Biol Med 138:104,920. https://doi.org/10.1016/j.compbiomed.2021.104920
61. García-Méndez S, de Arriba-Pérez F, Barros-Vila A et al (2022) Detection of temporality at discourse level on financial news by combining natural language processing and machine learning. Exp Syst Appl 197:116,648. https://doi.org/10.1016/j.eswa.2022.116648
62. Krippendorff K (2018) Content analysis: an introduction to its methodology. SAGE Publications
63. Sanchez-Gomez JM, Vega-Rodríguez MA, Pérez CJ (2018) Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. Knowl-Based Syst 159:1–8. https://doi.org/10.1016/j.knosys.2017.11.029
64. El-Kassas WS, Salama CR, Rafea AA, et al. (2020) EdgeSumm: graph-based framework for automatic text summarization. Inf Process Manag 57:102,264. https://doi.org/10.1016/j.ipm.2020.102264
65. Park H, Park T, Lee YS (2019) Partially collapsed Gibbs sampling for latent Dirichlet allocation. Exp Syst Appl 131:208–218. https://doi.org/10.1016/j.eswa.2019.04.028
66. Rash JA, Prkachin KM, Solomon PE et al (2019) Assessing the efficacy of a manual-based intervention for improving the detection of facial pain expression. Eur J Pain 23(5):1006–1019. https://doi.org/10.1002/ejp.1369
67. Seité S, Khammari A, Benzaquen M et al (2019) Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs. Exp Dermatol 28(11):1252–1257. https://doi.org/10.1111/exd.14022
68. Salminen J, Almerekhi H, Kamel AM et al (2019) Online hate ratings vary by extremes. In: Proceedings of the 2019, Conference on human information interaction and retrieval. Association for Computational Linguistics, pp 213–217, https://doi.org/10.1145/3295750.3298954
69. Kilicoglu H, Rosemblat G, Hoang L et al (2021) Toward assessing clinical trial publications for reporting transparency. J Biomed Inf 116:103,717–103,727. https://doi.org/10.1016/j.jbi.2021.103717
70. Gulden C, Kirchner M, Schüttler C et al (2019) Extractive summarization of clinical trial descriptions. Int J Med Inf 129:114–121. https://doi.org/10.1016/j.ijmedinf.2019.05.019
71. Hark C, Karcı A (2020) Karcı summarization: a simple and effective approach for automatic text summarization using Karcı entropy. Inf Process Manag 57(3):102,187. https://doi.org/10.1016/j.ipm.2019.102187
72. Alqaisi R, Ghanem W, Qaroush A (2020) Extractive multi-document Arabic text summarization using evolutionary multi-objective optimization with K-Medoid clustering. IEEE Access 8:228,206–228,224. https://doi.org/10.1109/ACCESS.2020.3046494

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Silvia García-Méndez received the Ph.D. degree in Information and Communication Technologies from University of Vigo in 2021. Since 2015, she has been working as a researcher with the Information Technologies Group at the University of Vigo. She is currently collaborating with foreign research centers as part of her postdoctoral stage. Her research interests include Natural Language Processing techniques and Machine Learning algorithms.

Francisco de Arriba-Pérez received the B.S. degree in telecommunication technologies engineering in 2013, the M.S. degree in telecommunication engineering in 2014, and the Ph.D. degree in 2019 from University of Vigo, Spain. He is currently a researcher in the Information Technologies Group at the University of Vigo, Spain. His research includes the development of Machine Learning solutions for different domains like finance and health.

Ana Barros-Vila received the B.S. degree in telecommunication technologies engineering from the University of Vigo, Spain, in 2020. Since 2019, she has been working as a researcher within the Information Technologies Group at the University of Vigo, Spain. Her research includes the development of classification systems using rule-based and Machine Learning approaches for financial knowledge extraction in Spanish and English languages.

Francisco J. González-Castaño received the B.S. degree from University of Santiago de Compostela, Spain, in 1990, and the Ph.D. degree from University of Vigo, Spain, in 1998. He is currently a full professor at the University of Vigo, Spain, where he leads the Information Technologies Group. He has authored over 100 papers in international journals in the fields of telecommunications and computer science and has participated in several relevant national and international projects. He holds three U.S. patents.

Enrique Costa-Montenegro received his Ph.D. degree from the University of Vigo, Spain, in 2007. He is currently an associate professor with the Department of Telematics Engineering at University of Vigo, Spain. His research interests include mobile services, natural language processing, wireless networks, multiagent systems, and recommendation technologies.

Affiliations

Silvia García-Méndez[1] · **Francisco de Arriba-Pérez**[1] · **Ana Barros-Vila**[1] · **Francisco J. González-Castaño**[1] · **Enrique Costa-Montenegro**[1]

Francisco de Arriba-Pérez
farriba@gti.uvigo.es

Ana Barros-Vila
abarros@gti.uvigo.es

Francisco J. González-Castaño
javier@gti.uvigo.es

Enrique Costa-Montenegro
kike@gti.uvigo.es

[1] Information Technologies Group, atlanTTic,
University of Vigo, E.I. Telecomunicación,
Campus Lagoas-Marcosende, Vigo, 36310, Spain