Expert Systems With Applications 197 (2022) 116648

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Detection of temporality at discourse level on financial news by combining Natural Language Processing and Machine Learning

Silvia García-Méndez *, Francisco de Arriba-Pérez, Ana Barros-Vila, Francisco J. González-Castaño

Information Technologies Group, atlanTTic, University of Vigo, E.I. Telecomunicación, Campus, 36310 Vigo, Spain

ARTICLE INFO

Keywords:
Computational Linguistics
Financial news
Knowledge extraction
Machine Learning
Natural Language Processing
Temporal analysis

ABSTRACT

Finance-related news such as Bloomberg News, CNN Business and Forbes are valuable sources of real data for market screening systems. In news, an expert shares opinions beyond plain technical analyses that include context such as political, sociological and cultural factors. In the same text, the expert often discusses the performance of different assets. Some key statements are mere descriptions of past events while others are predictions. Therefore, understanding the temporality of the key statements in a text is essential to separate context information from valuable predictions. We propose a novel system to detect the temporality of finance-related news at discourse level that combines Natural Language Processing and Machine Learning techniques, and exploits sophisticated features such as syntactic and semantic dependencies. More specifically, we seek to extract the dominant tenses of the main statements, which may be either explicit or implicit. We have tested our system on a labelled dataset of finance-related news annotated by researchers with knowledge in the field. Experimental results reveal a high detection precision compared to an alternative rule-based baseline approach. Ultimately, this research contributes to the state-of-the-art of market screening by identifying predictive knowledge for financial decision making.

1. Introduction

The field of Computational Linguistics has been prolific in theoretical work and practical solutions to real world challenges. This has become possible thanks to the increase in computing power, the development of enhanced architectures and Machine Learning models, and the advent of Web 4.0 online sources of relevant information. Natural Language Processing (NLP) is being broadly applied to these sources, from simple solutions using for example part-of-speech (POS) data, to more sophisticated designs exploiting syntactic and semantic dependencies.

Artificial understanding of a text to describe its meaning at discourse level, such as its temporal dimension, remains an open research question. Current work on knowledge extraction from text based on Artificial Intelligence (AI) techniques only covers limited aspects (Collobert et al., 2011).

Finance-related news from sources such as Bloomberg News, CNN Business and Forbes are valuable real data for market screening systems. News describe not only technical analyses of assets' performance but also their political, sociological and cultural contexts. Their content is informally organised around key statements that carry the main opinions of the author. Thus, the analysis of the temporal dimension of key statements becomes essential to differentiate predictions from the rest of the text, which also provides valuable information to interpret these statements, including their temporality.

In fact, every sentence in a text holds temporal knowledge (Zwaan, 1996). Temporal representation is an innate human capacity related to cognition and discourse processing, and there exist plenty of linguistic elements to express it (Demagny, 2012). More specifically, it involves certain linguistic markers (Evers-Vermeul et al., 2017) that typically combine lexical (temporal meaning of words), morphological (temporal features of languages such as tense), syntactic (position of time markers within the clause and their relation to other constituents) and pragmatic (discourse organisation and coherence through temporality) processings. Thus, temporal markers are not simply grammatical categories related to features like tense and aspect, but also temporal adverbial elements (for example: "a month ago", "today", "after", etc.) and complex verb structures such as phrasal verbs or compound verb phrases (e.g., "look forward", "begin to work").

* Corresponding author.
E-mail addresses: sgarcia@gti.uvigo.es (S. García-Méndez), farriba@gti.uvigo.es (F. de Arriba-Pérez), abarros@gti.uvigo.es (A. Barros-Vila), javier@det.uvigo.es (F.J. González-Castaño).

https://doi.org/10.1016/j.eswa.2022.116648
Received 23 November 2020; Received in revised form 27 September 2021; Accepted 4 February 2022
Available online 24 February 2022

Even though human beings excel in associative discourse inference, transferring this knowledge to computational systems is not straightforward (Pratt & Francez, 2001; Pratt-Hartmann, 2005). In this work, we address temporal analysis at discourse level with three different strategies. First, we apply clause segmentation to determine the continuity of tense in sequential clauses within the text, both across dependencies and by proximity. Second, we detect temporal modifiers, more specifically expressions that directly modify the temporal data of a clause, for example by altering verb tense. Finally, the positions of temporal references within the news are also considered, since authors tend to follow certain patterns, such as leaving predictions (future tenses) to the end. We apply these strategies along with Machine Learning techniques to determine the temporality of key opinions about assets in news. To the best of our knowledge, this work represents the first attempt to perform temporal analysis of finance news at discourse level. The rest of this article is organised as follows: Section 2 reviews related work on knowledge extraction and temporal analysis at discourse level in economics. Section 3 describes the model of the problem and our solution. Section 4 presents the experimental text corpus and the numerical tests that validate our approach. Finally, Section 5 concludes the paper.

2. Related work

Stock market screening with Data Mining has produced a significant body of research with successful outcomes that effectively support investment decision making (Alanyali et al., 2013; Day & Lee, 2016). Works such as Dimpfl and Jank (2016), Karabulut (2012), Nofer and Hinz (2015) have demonstrated the strong correlation between volatility and volume of search queries about stock market indexes.

In this field, NLP techniques have been successfully applied to noise removal and feature extraction (Fisher et al., 2016; Liu, 2015; Sun et al., 2014; Xing et al., 2018) from financial reports such as news (Alanyali et al., 2013; Atkins et al., 2018; Zhang & Skiena, 2010), micro-blogging comments (Fisher et al., 2016; Rickett, 2016; Sun et al., 2014; Wang, 2017; Xing et al., 2018) and social media (Ioans & Stoica, 2014; Sun et al., 2016). These techniques have been often combined with Machine Learning algorithms (Huang et al., 2012; Prollochs et al., 2015), which can be divided into supervised (based on manual annotations) (Alanyali et al., 2013; Prollochs et al., 2015) and unsupervised approaches (Huang et al., 2012; Prollochs et al., 2015).

Alanyali et al. (2013) demonstrated the relation between stock market events and public data in financial news. Atkins et al. (2018) applied Machine Learning models of Latent Dirichlet Allocation to predict stock market volatility. They concluded that the information extracted from news sources was more adequate than price variations for predicting the volatility of financial assets. Other authors have applied Deep Learning algorithms to financial news (Day & Lee, 2016; Vargas et al., 2017). Specifically, Day and Lee (2016) concluded that the particular typology of the source (news, micro-blogging comments, etc.) has strong influence in the resulting decision; while Vargas et al. (2017) proved that recurrent Neural Networks are better at capturing context information and modelling complex temporal characteristics for stock market forecasting.

Despite the several works on stock market knowledge extraction, the analysis of the temporal dimension deserves attention, since in most of the literature (Atkins et al., 2018; Rickett, 2016; Sun et al., 2016; Wang, 2017; Zhang & Skiena, 2010), temporality is exclusively determined from news, posts or comment timestamps. Early research (Schumann, 1987) examined the expression of temporality from different linguistic perspectives: morphology, semantics and pragmatics. Strongly aligned with our view, Gibbs (1998) highlighted the importance of analysing the time dimension from the customer perspective, i.e. the consumers' own understanding of time. Furthermore, Forray and Woodilla (2005) extracted time-related knowledge from journal titles. They paid attention to features such as punctuation, word choice, the use of academic terminology and keywords, as well as to context information.

Among recent research, there exist three general approaches to temporal analysis. First, in Kehler (2002), temporal relations are simply listed among other discourse relations. Conversely, in other works, different types of temporal relations constitute a separate class of discourse relation. This is the case of the Penn Discourse Treebank (Prasad et al., 2008) and the Rhetorical Structure Theory Discourse Treebank (Carlson et al., 2001). Finally, the third approach does not consider temporal order as a relational feature but as a pragmatic segment-specific feature (Sanders et al., 1993).

Considerable research on temporality has focused on verbs and, more specifically, on the analysis of verb tenses (Evers-Vermeul et al., 2017; Karapandza, 2016). Tense is a language feature that situates an event in time (Salaberry, 2003). Thus, it allows to contextualise the event within a temporal frame. In the specific case of English, it follows a rather clear scheme that simplifies the comprehension of temporal discourse organisation (Demagny, 2012). In other words, past, present and future tense morphologies often represent the corresponding times accurately.

In this paper, we also employ temporality at sentence level (the location of events on a timeline as indicated by past, present or future tenses), but we focus on the discourse level, that is, the relational organisation of the discourse due to the temporal ordering. As far as we know, no other state-of-the-art research has analysed the temporality of financial texts as expressed in natural language in the content itself.

3. System architecture

In this section we present a novel two-stage system to detect at discourse level the temporality of finance-related news, by combining NLP techniques, to extract sophisticated features like syntactic and semantic implicit dependencies, with a Machine Learning model. We seek to detect the general tense (past or future) when the author expresses an opinion about a stock asset, but we are specially interested in the future, which is strongly related to predictive analysis. To determine syntactic and semantic dependencies, we consider explicit mentions to stock market assets acting as subjects and objects in the sentences. Fig. 1 shows the general architecture of our system. In the next subsections we describe its modules.

3.1. Preprocessing

This module comprises the different techniques to remove unnecessary and redundant information from the input dataset.

First, before the replacement procedures, it is necessary to homogenise numerical data and dates in the experimental dataset. We rewrite them to a common format by applying regular expressions. Next, we replace all numerical values including percentages and dates with NUM, $PERC$ and $DATE$ tags, respectively.

Then, we use the Name Entity Classification (NEC) functionality from Freeling (Atserias et al., 2006; Padró & Stanilovsky, 2012) to detect proper names and locations. We double-check this detection with freely available lexica.[1,2] We also use freely available lexica to detect abbreviations.[3] At the end, these elements are replaced by tags $NAME$ (proper names), LOC (locations), and ABB (abbreviations).

[1] Available at https://names.mongabay.com/most_common_surnames.htm, November 2020.

[2] Available at https://datahub.io/core/world-cities#readme, November 2020.

[3] Available at https://www.gti.uvigo.es/index.php/en/resources/9-abbreviation-lexicon, November 2020.

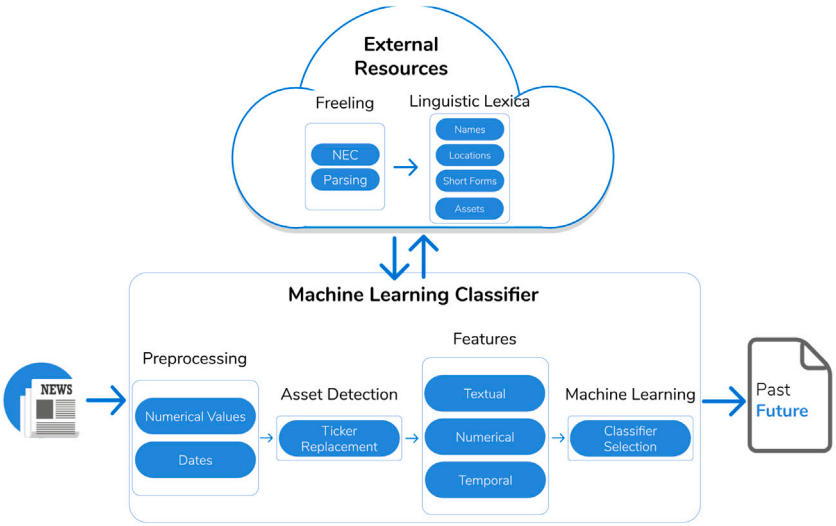

Fig. 1. System architecture.

Table 1
Example of news content before and after processing.

	News content
Before	If they could get the planes, **Airbus** and **Boeing** are sold out through **2023**. **On October 29**, **2018**, the **stock** dropped **6.6%** and recovered in three days.
After	If they could get the planes, **OTHER** and **TICKER** are sold out through **DATE**. On **DATE**, **DATE**, the **TICKER** dropped **NUM** and recovered in three days.

3.2. Asset detection

Notwithstanding the difficulty to extract stock market assets when neither exact names nor pronouns are used, we detect referential elements, such as nouns (e.g. "company", "enterprise", "stock", "investment", "opportunity", "share", "manufacturer", etc.). More specifically, if the referential element appears after one or several tickers (within the same sentence), it is replaced by the last ticker; otherwise, it is replaced by the last ticker of the previous sentence (if any).

We use tag $TICKER$ to replace the main asset in the news, as indicated by the annotators, and tag $OTHER$ for any other assets. All these other assets in each news piece are extracted using a finance lexicon.[4] Table 1 shows a complete example of asset detection and replacement (note that it also exemplifies numerical data and date detection and replacement, see Section 3.1).

3.3. Features and Machine Learning analysis

Our system employs ZeroR, Decision Tree (DT), Random Forest (RF), Linear Support Vector Classification (SVC) and Neural Network (NN) algorithm implementations from the Scikit-Learn Python library.[5] We included ZeroR as a reference for the performance of the rule-based baseline classifier due to its simplicity and low computational cost (Fazayeli et al., 2019).

Table 2 shows the training and testing complexity of the Machine Learning algorithms we used in our analysis for c target classes, f features, i algorithm instances (where applicable) and s dataset samples. For the specific case of the NN algorithm, m represents the number of neurons, and l its layers. The DT and RF algorithms have logarithmic training complexity (Hassine et al., 2019; Witten et al., 2016), that is, less than SVC (Vapnik, 2000) (which, however, has a low classification response time compared to other alternatives when using a linear kernel, as in our work). At the end, NN has the highest training and testing complexity (Han et al., 2012; Witten et al., 2016).

Table 2
Machine Learning training and testing complexity order.

Classifier	Train complexity	Test complexity
DT	$O(n \cdot \log(n) \cdot f)$	O(depth of the tree)
RF	$O(n \cdot \log(n) \cdot f \cdot i)$	O(depth of the tree$\cdot i$)
SVC	$O(s^2)$	$O(f)$
NN	$O(m \cdot f \cdot s \cdot l)$	$O(m \cdot f \cdot l)$

Fig. 2 shows the Machine Learning process. There are $f = 30$ features in our system as indicated in Table 3, divided into textual, numerical and temporal features.

Before computing the textual features, the text is preprocessed to replace capital by lowercase characters and remove punctuation marks (commas, colons, brackets, hyphens, exclamation and interrogation marks), accents and apostrophes, as well as characters @ and #. This is applied to the textual content excluding the tags identified in prior preprocessing and asset detection stages (Sections 3.1 and 3.2, respectively).

At the end, the resulting textual features are n-grams. More specifically, char-grams, word tokens (character n-grams only from text inside word boundaries) and word-grams. We explain the n-gram feature selection in Section 4.3.

The numerical features are basically the numerical values in the news content. There are two such features: the amount of numerical values excluding percentages and the amount of percentages in the text. Note that we discard explicit dates, although it would be possible to determine if such dates refer to the future or the past if the news publication date is available.

Regarding temporal features, we use the dependency parsing by Freeling to extract the verbs related to the tickers and their tenses. In addition, we analyse the proximity between a verb and a ticker within

[4] Available at https://www.gti.uvigo.es/index.php/en/resources/10-financial-lexicon, November 2020.

[5] Available at https://scikit-learn.org/stable/supervised_learning.html#supervised-learning, November 2020.

Table 3
Features of the temporality detection system.

Type	Feature name	Description
Textual	CHAR_GRAMS	Character *n*-grams from the news content.
	WORD_TOKENS	Character *n*-grams only from text inside word boundaries.
	WORD_GRAMS	Word *n*-grams from the news content.
Numerical	NUM	Amount of numerical values (excluding percentages) in the news content.
	PERC	Amount of percentages in the news content.
Temporal	PRS_DEP_SUB	Amount of verbs in present tense from the dependency analysis when the ticker acts as subject.
	PST_DEP_SUB	Amount of verbs in past tense from the dependency analysis when the ticker acts as subject.
	FUT_DEP_SUB	Amount of verbs in future tense from the dependency analysis when the ticker acts as subject.
	GLOBAL_DEP_SUB	Global temporality by majority voting from the dependency analysis when the ticker acts as subject.
	PRS_DEP_SUB_OBJ	Amount of verbs in present tense from the dependency analysis when the ticker acts either as subject or object.
	PST_DEP_SUB_OBJ	Amount of verbs in past tense from the dependency analysis when the ticker acts either as subject or object.
	FUT_DEP_SUB_OBJ	Amount of verbs in future tense from the dependency analysis when the ticker acts either as subject or object.
	GLOBAL_DEP_SUB_OBJ	Global temporality by majority voting from the dependency analysis when the ticker acts either as subject or object.
	PRS_PROX_SUB	Amount of verbs in present tense from the proximity analysis when the ticker acts as subject.
	PST_PROX_SUB	Amount of verbs in past tense from the proximity analysis when the ticker acts as subject.
	FUT_PROX_SUB	Amount of verbs in future tense from the proximity analysis when the ticker acts as subject.
	GLOBAL_PROX_SUB	Global temporality by majority voting from the proximity analysis when the ticker acts as subject.
	PRS_PROX_SUB_OBJ	Amount of verbs in present tense from the proximity analysis when the ticker acts either as subject or object.
	PST_PROX_SUB_OBJ	Amount of verbs in past tense from the proximity analysis when the ticker acts either as subject or object.
	FUT_PROX_SUB_OBJ	Amount of verbs in future tense from the proximity analysis when the ticker acts either as subject or object.
	GLOBAL_PROX_SUB_OBJ	Global temporality by majority voting from the proximity analysis when the ticker acts either as subject or object.
	PRS_INITIAL	Amount of verbs in present tense in the initial third of the news.
	PST_INITIAL	Amount of verbs in past tense in the initial third of the news.
	FUT_INITIAL	Amount of verbs in future tense in the initial third of the news.
	PRS_MEDIUM	Amount of verbs in present tense in the middle third of the news.
	PST_MEDIUM	Amount of verbs in past tense in the middle third of the news.
	FUT_MEDIUM	Amount of verbs in future tense in the middle third of the news.
	PRS_FINAL	Amount of verbs in present tense in the final third of the news.
	PST_FINAL	Amount of verbs in past tense in the final third of the news.
	FUT_FINAL	Amount of verbs in future tense in the final third of the news.

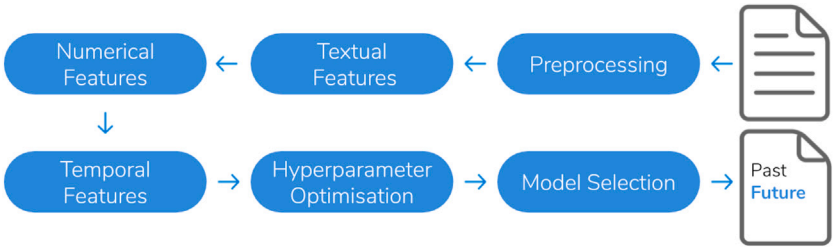

Fig. 2. Flow diagram of Machine Learning analysis.

a clause in forward and backward mode. These analyses are applied regardless of the role of each ticker instance (subject or object). Taking the text *And second, will TICKER be allowed to execute its plan in full? Cost-cutting should make TICKER a leaner[...]* as an example, note that the first $TICKER$ tag acts as a subject while Freeling considers the second tag an object.

Observe that, in the dependency analysis, in case the ticker is a modifier of the subject (e.g. "Intel stock", where "stock" is the subject and "Intel" its modifier), and therefore it is not directly related to any verbs, we also consider the verbs that are linked to the subject. In the proximity analysis, given a ticker and the immediately preceding and posterior verbs within the clause, the nearest verb in number of intermediate words is selected. Table 4 shows an example of news entry after applying dependency and proximity analysis. We highlight the tickers and the verbs that are related to them. Note that in both analyses the ticker may act as subject or object. Finally, we also consider the positions of the verbs in the discourse, by dividing the news pieces into three roughly equal parts in number of sentences.

Summing up, we propose three temporal features (number of past, present and future tenses, when the ticker acts as subject and when it acts either as subject or object) for the dependency analysis, the same for the proximity analysis, one feature with the global temporality for the dependency analysis and another similar feature for proximity analysis (both by majority voting). There are three additional features per discourse partition (initial, middle, final) that simply count the respective amounts of verb tenses. This completes 25 temporal features.

In case of tie between past and future tenses, the latter prevails. Even though the system only predicts past and future temporality, it also takes present tenses as input. Note that we apply a hyperparameter optimisation to the set of classifiers. This allows us to select the best model based on its performance (see Section 4.4).

4. Experimental results and discussion

In this section we first present some preliminary results using a rule-based approach as a baseline, and then evaluate our system. A rule-based reference is common practice in the literature when similar competitors are missing (Araque et al., 2017; Cronin et al., 2017; Jørgensen & Igel, 2021; Kim et al., 2020; Singh Chauhan et al., 2020). All experiments were performed on a computer with the following hardware 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

Table 4
Example of news entry after applying dependency and proximity analysis when the ticker acts as subject or object.

Dependency analysis	Proximity analysis
@TICKER@ is lagging on its competitors. Make no mistake, @TICKER@ is going to have to fix @TICKER@ and @TICKER@ will take many, many, many years, @NAME@ Mosesmann, a technology analyst at @NAME@ Securities, told CNBC in an interview. Couple that issue with an ongoing search to replace former chief executive officer @NAME@ Krzanich, and **@TICKER@ has** a lot on its plate heading into the last quarter of @DATE@ and beyond.	**@TICKER@ is lagging** on its competitors. Make no mistake, **@TICKER@ is going to have to fix** @TICKER@ and **@TICKER@ will take** many, many, many years, @NAME@ Mosesmann, a technology analyst at @NAME@ Securities, told CNBC in an interview. Couple that issue with an ongoing search to replace former chief executive officer @NAME@ Krzanich, and **@TICKER@ has** a lot on its plate heading into the last quarter of @DATE@ and beyond.

Table 5
Example of news entry in the dataset.

Text	Ticker	Source	Temporality
Pros and Cons to Buying Intel Stock Intel is lagging on its competitors. Can its stock price holding up under the pressure? "Make no mistake, Intel is going to have to fix this and it will take many, many, many years.", Hans Mosesmann, a technology analyst at Rosenblatt Securities, told CNBC in an interview. "Their process technology disadvantage, which I think is broken, will take five, six, seven years. I don't think that business model works by them being behind by a year or two in terms of process technology." Couple that issue with an ongoing search to replace former chief executive officer Brian Krzanich, and Intel has a lot on its plate heading into the last quarter of 2019 and beyond.	Intel	US News	Future

Table 6
Distribution of entries in the dataset by category.

Temporal tag	Number of entries	Avg. length in sentences	Avg. length in words
Past	365	21.67 ± 13.89	351.57 ± 225.82
Future	235	18.43 ± 12.41	303.40 ± 196.36
Total	600	20.05 ± 13.15	327.49 ± 211.09

4.1. Dataset

Our experimental dataset[6] is composed of $s = 600$ news pieces with the following information: identifier, content, ticker (main stock market asset discussed), source and temporality (past and future tags of key opinions in the text). The sources of the news include prestigious journals such as The New York Times and stock messaging boards such as Bloomberg. Table 5 shows an example of news entry in the dataset. Table 6 shows the distribution of its entries. The dataset is comparable to other datasets in previous research on knowledge extraction (García et al., 2018; Li et al., 2018).

4.2. Rule-based baseline results

Before applying our Machine Learning approach, we tested a baseline system based on syntactic and semantic rules to detect temporality from news. First, we created a financial semantic tree using the data provided by Multilingual Central Repository[7] (MCR) (González-Agirre et al., 2012). The hierarchical elements of interest were: "commerce", "enterprise", "finance", "banking", "exchange", "money", "insurance", "tax" and "industry". Then we applied the semantic tree to analyse the news content at word level. The analysis was divided into two parts:

- We checked if any word in the news title was included in our financial semantic tree. Then, we created an extractive summary of the news content discarding those sentences that did not include any word belonging to the aforementioned financial semantic categories. The sentences that included the main ticker were kept. In the event that no word in the title was related to finance, based on the information extracted from the semantic tree, we created the extractive summary by applying a TF-IDF approach with a relevance threshold of 0.75 inspired by similar works in the literature (Rabelo et al., 2020; Suleman & Korkontzelos, 2021; Xiao & Tong, 2021).
- By considering the dependency parsing analysis and using the extractive summary from the previous step, each sentence was analysed with Freeling. First, we considered the verbs with a syntactic relation with a ticker or a referential element acting as the main element of the clause. If no such verbs were detected, the closest previous or posterior verb to a ticker or referential element was taken into account. If no verbs were detected, we divided the sentences by commas and explored their parts separately.

With a slight abuse of notation, let us refer to the number of instances of the respective verb tenses as "past", "present" and "future". We applied the following rules when either future or past > 0 to decide the temporality of the news:

1. Future, if one of the following criteria is met:
   (a) future ≥ past
   (b) past> 1 and present + future > past
   (c) present ≥ 3 × past
2. Otherwise, past

These rules have resulted from experimental tests with good performance. We performed a combinatorial search over configurable parameter ranges for the number of past, present and future references used as estimators.

Also based on experimental tests, where neither past nor future temporality was detected (when both future and past = 0), the system

[6] The experimental dataset will be made available to other researchers on request.

[7] A lexical database that integrates the Spanish WordNet into the EuroWordNet framework, available at http://adimen.si.ehu.es/web/MCR, November 2020.

Table 7
Macro performance of the rule-based baseline approach.

Precision	Recall	Accuracy
73.48%	74.47%	74.57%

Table 8
Macro performance and training and testing times using selected textual features.

Classifier	Precision	Recall	Accuracy	Train (s)	Test (s)
ZeroR	30.42	50.00	60.83	<0.01	<0.01
DT	65.82%	66.06%	67.17%	0.85	0.54
RF	77.48%	76.61%	78.17%	7.92	0.61
SVC	**81.70%**	**82.12%**	**82.50%**	2.83	0.55
NN	79.38%	79.33%	79.67%	380.46	0.89

Table 9
Macro performance and training and testing times using selected textual features plus all numerical and temporal features.

Classifier	Precision	Recall	Accuracy	Train (s)	Test (s)
ZeroR	30.42	50.00	60.83	<0.01	<0.01
DT	67.12%	67.29%	68.33%	1.23	0.58
RF	79.71%	78.59%	80.17%	5.81	0.64
SVC	**83.08%**	**82.38%**	**83.50%**	3.60	0.60
NN	82.68%	82.34%	83.17%	505.28	0.94

considers that, if a present tense is followed by a number, it is a reference to the past (else, to the future). To explain the first condition take as an example the sentence "On October 29, 2018, the stock dropped 6.6%" in Table 1, where the main verb is in past tense. In financial jargon this kind of statement is sometimes expressed in present tense. However, an explicit quantity is strongly indicative that the event has already occurred, unless there is an explicit reference to the future.

We obtained near-75% accuracy using this rule-based baseline approach (see Table 7). We considered these results promising but they motivated us to pursue more sophisticated Machine Learning techniques to extract temporality from financial news content with higher performance.

4.3. Feature tuning and training

To create the Machine Learning model, we chose the most adequate features from our annotated dataset.

For the generation of the *n*-grams (char-grams, word tokens and word-grams), we used *GridSearchCV*[8] from the Scikit-Learn Python library, which is an exhaustive search of an estimator in specified parameter ranges. We selected wide ranges in *CountVectorizer*[9] (as shown in Listing 1). As a result, we obtained the following optimal parameters: `max_df` = 0.30, `min_df` = 0, `ngram_range` = (2,4) and `max_features` = 10000.

Listing 1: Configuration parameters for the generation of *n*-grams.

```
max_df: (0.3,0.35,0.4,0.5,0.7,0.8,1)
min_df: (0, 0.002, 0.005,0.008,0.01)
ngram_range: ((1, 1),(1, 2),(1, 3),(1, 4),(2, 4))
max_features: (10000,20000,30000,None)
```

At the end, we obtained 30,000 features only for the *n*-grams (10,000 per type, char-grams, word-grams and word tokens). This huge set of features could lead to computationally unfeasible models and excessively long prediction times. Indeed, redundant features or poorly correlated features with the target predictor would cause unacceptable system performance. Therefore, we applied an attribute selector to extract the most relevant features with 10-fold cross-validation. For this purpose we chose the *SelectPercentile*[10] method from the Scikit-Learn Python library, as it outperformed other alternatives[10] (*SelectFrom Model*, *SelectKBest*, and *RFECV*). This method selects features according to a percentile of the highest scores. We set Chi^2 as the score function and an 80th percentile threshold. With this approach, *n*-gram features dropped to 24,000.

Once these relevant *n*-gram features were selected, we analysed the temporal and numerical features, and we selected the best ones with a combinatorial analysis (one-vs-rest analysis for all 27 features) with 10-fold cross-validation (see Section 4.4). For all Machine Learning analyses, we applied a hyperparameter optimisation using *GridSearchCV* over the classification models with 10-fold cross-validation.

Table 10
Performance by class using selected textual features.

Classifier	Precision		Recall	
	Past	Future	Past	Future
SVC	**87.12%**	76.28%	83.81%	**80.42%**
NN	81.41%	**77.21%**	**86.28%**	67.63%

Table 11
Performance by class using selected textual features plus all numerical and temporal features.

Classifier	Precision		Recall	
	Past	Future	Past	Future
SVC	85.89%	**80.28%**	**87.36%**	77.39%
NN	**86.47%**	78.88%	85.98%	**78.71%**

4.4. Machine learning results

We performed three experiments: only using the selected *n*-grams (24,000 features), using the selected *n*-grams plus the two numerical features and all temporal features (25 features) and, finally, using the selected *n*-grams plus the most relevant numerical and temporal features from a combinatorial selection that we explain later in this section. In each of these three experiments we tested all Machine Learning classifiers in the system with 10-fold cross-validation on the experimental dataset. Since our main objective is detecting predictions of future events in news, we paid special attention to the results for this class.

Table 8 shows the results obtained with *n*-grams. The best classification model was SVC followed by NN. However, the latter is computationally expensive due to the configuration time of the different hidden layers for each feature in the model. By comparing these results with those of the second experiment (Table 9), we observe a ~4% improvement in accuracy for the NN classifier thanks to the numerical and temporal features. In any case, the improvement of the SVC classifier in Table 9 over the rule-based baseline was at least 8% for all metrics.

The effect of numerical and temporal features became more apparent when we checked the behaviour by class. Table 10 shows the results of the first experiment in that case. Note that precision and recall were very asymmetric between past and future (~10% precision asymmetry with the SVC classifier, ~19% recall asymmetry with the NN classifier). In addition, the precision of both classifiers was barely above 75% for future.

The introduction of temporal features solves these performance issues (asymmetric precision and levels under 80%) to a great extent.

[8] Available at https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html, November 2020.

[9] Available at https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html, November 2020.

[10] Available at https://scikit-learn.org/stable/modules/feature_selection.html, November 2020.

Table 12
Macro performance and training and testing times using selected textual features and most relevant temporal features from the combinatorial analysis.

Classifier	Precision	Recall	Accuracy	Train (s)	Test (s)
SVC	**83.88%**	**83.59%**	**84.33%**	3.47	0.60
NN	79.80%	79.22%	80.00%	571.25	0.94

Table 13
Performance by class using selected textual features and most relevant temporal features from the combinatorial analysis, SVC classifier.

Classifier	Precision		Recall	
	Past	Future	Past	Future
SVC	87.54%	80.21%	86.79%	80.40%

Accordingly, Table 11 shows a 4% improvement in precision for the future class with a SVC classifier. For the NN classifier, precision improvements were ~5% and ~2% for past and future, respectively, although the precision for the future class only reached ~ 79%. Note however the high improvement in recall with the NN classifier for the future class, which exceeded 10%.

We then focused on the SVC classifier, whose recall for the future class was still under 80% but presented the best overall results in terms of precision, recall and accuracy. Here we introduced the previously mentioned selection of the most relevant numerical and temporal features with a combinatorial analysis with 10-fold cross-validation. We chose 11 such features due to the precision they attained: number of present and future verbs from the dependency analysis with the ticker acting as subject, number of present verbs from the dependency parsing with the ticker acting either as subject or object, number of present and future verbs from the proximity analysis with the ticker acting as subject, number of present and future verbs from the proximity analysis with the ticker acting either as subject or object, global temporality from the proximity analysis with the ticker acting either as subject or object, number of future verbs in the second part of the news piece and number of present and future verbs in the third part. Only the features that produced precision symmetry between classes were kept, and both numerical features were discarded.

Table 12 shows that, with this second selection, we attained well over 80% precision and recall performance with the SVC classifier, which takes considerably less time to train than the NN. Furthermore, Table 13 shows the precision and recall of the SVC classifier by class. Note that all metrics exceeded 80% as pursued, a level that, compared to other Machine Learning financial applications in the literature (Atkins et al., 2018; De Arriba-Pérez et al., 2020; Dridi et al., 2019; Zhu et al., 2017, 2019), is similar and even superior.

4.5. Application use case

A direct use case of our system would be a financial mobile application ("app") for investors, which would automatically process news sources. The app would translate this vast amount of public information into key indicators on assets, including overall temporality of key statements. Fig. 3 shows a possible interface of this app, where future statements are marked in green and assets are highlighted in pink. The confidence in the classification is also presented to the users at the bottom.

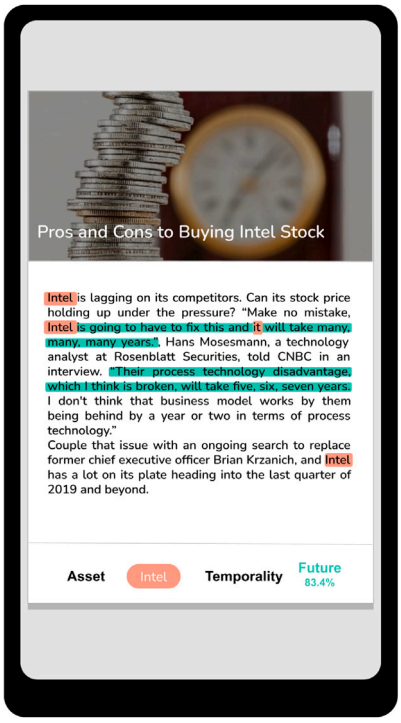

Fig. 3. Example of temporality detection system for financial news integrated into a mobile application.

5. Conclusions

Motivated by the amount of publicly available data about stock markets in financial news, we propose a novel system to detect the temporal dimension of the key statements in a text at discourse level, which relies on a combination of NLP and Machine Learning techniques.

For that purpose, we departed from a manually annotated dataset of 600 financial news. In addition to textual features (*n*-grams) and numerical features, we applied different analyses to the text to obtain temporal features: dependency parsing, verb extraction by proximity and verb analysis by text partitions (as an approximation to discourse patterns in financial news). Thus, the more sophisticated features required syntactic and dependency parsing NLP techniques. These features were essential for the Machine Learning algorithms to achieve precision and recall above 80%, both globally and by class.

Since we are not aware of any previous research on this same problem, we checked our approach against a rule-based baseline based on the temporal features. The Machine Learning approach was significantly better (between 8% and 10% improvement in all metrics).

As future work we plan to design a multilingual version of our architecture to also cover financial news in Spanish.

CRediT authorship contribution statement

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

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by Xunta de Galicia, Spain grants GRC2018/053 and ED341D-R2016/012.

References

Alanyali, M., Moat, H. S., & Preis, T. (2013). Quantifying the relationship between financial news and the stock market. *Scientific Reports*, *3*(1), 3578. http://dx.doi.org/10.1038/srep03578.

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, *77*, 236–246. http://dx.doi.org/10.1016/j.eswa.2017.02.002.

Atkins, A., Niranjan, M., & Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, *4*(2), 120–137. http://dx.doi.org/10.1016/j.jfds.2018.02.002.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., & Padró, M. (2006). FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In *Proceedings of the international conference on language resources and evaluation* (pp. 2281–2286). European Language Resources Association.

Carlson, L., Marcu, D., & Okurowski, M. E. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the SIGdial workshop on discourse and dialogue (Vol. 16)* (pp. 1–10). Association for Computational Linguistics, http://dx.doi.org/10.3115/1118078.1118083.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *2*(8), 2493–2537, arXiv:1103.0398.

Cronin, R. M., Fabbri, D., Denny, J. C., Rosenbloom, S. T., & Jackson, G. P. (2017). A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics*, *105*, 110–120. http://dx.doi.org/10.1016/j.ijmedinf.2017.06.004.

Day, M.-Y., & Lee, C.-C. (2016). Deep learning for financial sentiment analysis on finance news providers. In *Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 1127–1134). IEEE, http://dx.doi.org/10.1109/ASONAM.2016.7752381.

De Arriba-Pérez, F., García-Méndez, S., Regueiro-Janeiro, J. A., & González-Castaño, F. J. (2020). Detection of financial opportunities in micro-blogging data with a stacked classification system. *IEEE Access*, *8*, 215679–215690. http://dx.doi.org/10.1109/ACCESS.2020.3041084.

Demagny, A. C. (2012). Paths in L2 acquisition: The expression of temporality in spatially oriented narration. In *Comparative perspectives on language acquisition: A tribute to clive perdue (Vol. 61)* (pp. 482–501). Multilingual Matters Publisher.

Dimpfl, T., & Jank, S. (2016). Can internet search queries help to predict stock market volatility? *European Financial Management*, *22*(2), 171–192. http://dx.doi.org/10.1111/eufm.12058.

Dridi, A., Atzeni, M., & Reforgiato Recupero, D. (2019). FineNews: fine-grained semantic sentiment analysis on financial microblogs and news. *International Journal of Machine Learning and Cybernetics*, *10*(8), 2199–2207. http://dx.doi.org/10.1007/s13042-018-0805-x.

Evers-Vermeul, J., Hoek, J., & Scholman, M. C. (2017). On temporality in discourse annotation: Theoretical and practical considerations. *Dialogue and Discourse*, *8*(2), 1–20. http://dx.doi.org/10.5087/dad.2017.201.

Fazayeli, H., Syed-Mohamad, S. M., & Md Akhir, N. S. (2019). Towards auto-labelling issue reports for pull-based software development using text mining approach. *Procedia Computer Science*, *161*, 585–592. http://dx.doi.org/10.1016/j.procs.2019.11.160.

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 157–214. http://dx.doi.org/10.1002/isaf.1386.

Forray, J. M., & Woodilla, J. (2005). Artefacts of management academe. *Time & Society*, *14*(2–3), 323–339. http://dx.doi.org/10.1177/0961463X05055142.

García, S., Zhang, Z.-L., Altalhi, A., Alshomrani, S., & Herrera, F. (2018). Dynamic ensemble selection for multi-class imbalanced datasets. *Information Sciences*, *445-446*, 22–37. http://dx.doi.org/10.1016/j.ins.2018.03.002.

Gibbs, P. (1998). Time, temporality and consumer behaviour. *European Journal of Marketing*, *32*(11/12), 993–1007. http://dx.doi.org/10.1108/03090569810243622.

González-Agirre, A., Laparra, E., & Rigau, G. (2012). Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the global wordnet conference* (pp. 118–125). Tribun EU.

Han, J., Kamber, M., & Pei, J. (2012). Data mining. *Data mining: Concepts and techniques* (p. 740). Elsevier, http://dx.doi.org/10.1016/C2009-0-61819-5.

Hassine, K., Erbad, A., & Hamila, R. (2019). Important complexity reduction of random forest in multi-classification problem. In *International wireless communications & mobile computing conference* (pp. 226–231). IEEE, http://dx.doi.org/10.1109/IWCMC.2019.8766544.

Huang, S.-Y., Tsaih, R.-H., & Lin, W.-Y. (2012). Unsupervised neural networks approach for understanding fraudulent financial reporting. *Industrial Management & Data Systems*, *112*(2), 224–244. http://dx.doi.org/10.1108/02635571211204272.

Ioans, E., & Stoica, I. (2014). Social media and its impact on consumers behavior. *International Journal of Economic Practices and Theories*, *4*(2), 295–303.

Jørgensen, R. K., & Igel, C. (2021). Machine learning for financial transaction classification across companies using character-level word embeddings of text fields. *Intelligent Systems in Accounting, Finance and Management*, 1–14. http://dx.doi.org/10.1002/isaf.1500.

Karabulut, Y. (2012). Can facebook predict stock market activity? *SSRN Electronic Journal*, 1–58. http://dx.doi.org/10.2139/ssrn.2017099.

Karapandza, R. (2016). Stock returns and future tense language in 10-K reports. *Journal of Banking & Finance*, *71*, 50–61. http://dx.doi.org/10.1016/j.jbankfin.2016.04.025.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar* (p. 231). CSLI Publishers.

Kim, A., Yang, Y., Lessmann, S., Ma, T., Sung, M.-C., & Johnson, J. (2020). Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting. *European Journal of Operational Research*, *283*(1), 217–234. http://dx.doi.org/10.1016/j.ejor.2019.11.007, arXiv:1812.06175.

Li, Y., Pan, Q., Wang, S., Yang, T., & Cambria, E. (2018). A generative model for category text generation. *Information Sciences*, *450*, 301–315. http://dx.doi.org/10.1016/j.ins.2018.03.050.

Liu, B. (2015). *Sentiment analysis. mining opinions, sentiments, and emotions* (p. 383). Cambridge University Press, http://dx.doi.org/10.1017/CBO9781139084789.

Nofer, M., & Hinz, O. (2015). Using twitter to predict the stock market. *Business & Information Systems Engineering*, *57*(4), 229–242. http://dx.doi.org/10.1007/s12599-015-0390-4.

Padró, L., & Stanilovsky, E. (2012). Freeling 3.0: towards wider multilinguality. *Proceedings of the international conference on language resources and evaluation*, 2473–2479.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The penn discourse treebank 2.0. In *Proceedings of the international conference on language resources and evaluation* (pp. 2961–2968). European Language Resources Association.

Pratt, I., & Francez, N. (2001). Temporal prepositions and temporal generalized quantifiers. *Linguistics and Philosophy*, *24*(2), 187–222. http://dx.doi.org/10.1023/A:1005632801858.

Pratt-Hartmann, I. (2005). Temporal prepositions and their logic. *Artificial Intelligence*, *166*(1–2), 1–36. http://dx.doi.org/10.1016/j.artint.2005.04.003.

Prollochs, N., Feuerriegel, S., & Neumann, D. (2015). Enhancing sentiment analysis of financial news by detecting negation scopes. In *Proceedings of the Hawaii international conference on system sciences* (pp. 959–968). IEEE, http://dx.doi.org/10.1109/HICSS.2015.119.

Rabelo, J., Kim, M.-Y., Goebel, R., Yoshioka, M., Kano, Y., & Satoh, K. (2020). A summary of the COLIEE 2019 competition. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 34–49). Springer, http://dx.doi.org/10.1007/978-3-030-58790-1_3.

Rickett, L. K. (2016). Do financial blogs serve an infomediary role in capital markets? *American Journal of Business*, *31*(1), 17–40. http://dx.doi.org/10.1108/AJB-08-2015-0024.

Salaberry, R. (2003). Tense aspect in verbal morphology. *Hispania*, *86*(3), 559. http://dx.doi.org/10.2307/20062909.

Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1993). Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics*, *4*(2), 93–134. http://dx.doi.org/10.1515/cogl.1993.4.2.93.

Schumann, J. H. (1987). The expression of temporality in basilang speech. *Studies in Second Language Acquisition*, *9*(1), 21–41. http://dx.doi.org/10.1017/S0272263100006495.

Singh Chauhan, G., Kumar Meena, Y., Gopalani, D., & Nahta, R. (2020). A two-step hybrid unsupervised model with attention mechanism for aspect extraction. *Expert Systems with Applications*, *161*, 113673. http://dx.doi.org/10.1016/j.eswa.2020.113673.

Suleman, R. M., & Korkontzelos, I. (2021). Extending latent semantic analysis to manage its syntactic blindness. *Expert Systems with Applications*, *165*, 114130. http://dx.doi.org/10.1016/j.eswa.2020.114130.

Sun, F., Belatreche, A., Coleman, S., McGinnity, T. M., & Li, Y. (2014). Pre-processing online financial text for sentiment classification: a natural language processing approach. In *Proceedings of the IEEE conference on computational intelligence for financial engineering & economics* (pp. 122–129). IEEE, http://dx.doi.org/10.1109/CIFEr.2014.6924063.

Sun, A., Lachanski, M., & Fabozzi, F. J. (2016). Trade the tweet: social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*, *48*, 272–281. http://dx.doi.org/10.1016/j.irfa.2016.10.009.

Vapnik, V. N. (2000). *The nature of statistical learning theory*. Springer New York, http://dx.doi.org/10.1007/978-1-4757-3264-1.

Vargas, M. R., de Lima, B. S. L. P., & Evsukoff, A. G. (2017). Deep learning for stock market prediction from financial news articles. In *Proceedings of the IEEE international conference on computational intelligence and virtual environments for measurement systems and applications* (pp. 60–65). IEEE, http://dx.doi.org/10.1109/CIVEMSA.2017.7995302.

Wang, Y. (2017). Stock market forecasting with financial micro-blog based on sentiment and time series analysis. *Journal of Shanghai Jiaotong University (Science)*, *22*(2), 173–179. http://dx.doi.org/10.1007/s12204-017-1818-4.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers, http://dx.doi.org/10.1016/c2009-0-19715-5.

Xiao, S., & Tong, W. (2021). Prediction of user consumption behavior data based on the combined model of TF-IDF and logistic regression. *Journal of Physics: Conference Series*, *1757*(1), 012089. http://dx.doi.org/10.1088/1742-6596/1757/1/012089.

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, *50*(1), 49–73. http://dx.doi.org/10.1007/s10462-017-9588-9.

Zhang, W., & Skiena, S. (2010). Trading strategies to exploit blog and news sentiment. In *Proceedings of the international AAAI conference on weblogs and social media* (pp. 376–378). Association for the Advancement of Artificial Intelligence.

Zhu, Y., Xie, C., Wang, G.-J., & Yan, X.-G. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications*, *28*(S1), 41–50. http://dx.doi.org/10.1007/s00521-016-2304-x.

Zhu, Y., Zhou, L., Xie, C., Wang, G.-J., & Nguyen, T. V. (2019). Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics*, *211*, 22–33. http://dx.doi.org/10.1016/j.ijpe.2019.01.032.

Zwaan, R. A. (1996). Processing narrative time shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1196–1207. http://dx.doi.org/10.1037/0278-7393.22.5.1196.