information

MDPI

Article

Finding Evidence of Fraudster Companies in the CEO's Letter to Shareholders with Sentiment Analysis

Núria Bel [1,*], Gabriel Bracons [2] and Sophia Anderberg [1]

[1] Departament de Traducció i Ciències del Llenguatge, Universitat Pompeu Fabra, E-08018 Barcelona, Spain; sof.anderberg@gmail.com
[2] School of Economics and Finance, Queen Mary University of London, London E1 4NS, UK; g.braconsfont@qmul.ac.uk
* Correspondence: nuria.bel@upf.edu

Abstract: The goal of our research was to assess whether the observation about deceptive texts having a lower positive tone than truthful ones in terms of sentiment could become operative and be used for building a classifier in the particular case of fraudster's letters written in Spanish. The data were the letters that CEOs address to company shareholders in their annual financial reports, and the task was to identify the letters of companies that committed financial misconduct or fraud. This case was challenging for two reasons: first, most of the research worked with spontaneous written or spoken texts, while these letters did not; second, most of the research in this area worked on English texts, while we validated the linguistic cues found as evidence of deception for Spanish texts. The results of our research confirm that an SVM trained with a bag-of-words model of frequent adjectives can achieve 81% accuracy because these adjectives bring the information about which positive or negative tone and which word combinations in a text turn out to be a characteristic of fraudster's texts.

Keywords: fraud identification; text classification; deceptive text; sentiment analysis; SVM

Citation: Bel, N.; Bracons, G.; Anderberg, S. Finding Evidence of Fraudster Companies in the CEO's Letter to Shareholders with Sentiment Analysis. *Information* **2021**, *12*, 307. https://doi.org/10.3390/info12080307

Academic Editors: Salud María Jiménez-Zafra and Miguel Ángel García Cumbreras

Received: 27 June 2021
Accepted: 28 July 2021
Published: 30 July 2021

1. Introduction

Sentiment analysis (SA) is the area of natural language processing that is focused on identifying subjective information, such as the polarity (positive, negative, or neutral) of the writer's opinion and other manifestations of people's emotions as expressed in texts [1]. In this paper, we present the results of the application of SA to letters that CEOs address to Spanish company shareholders in annual financial reports. In these letters, management decisions are explained and justified. The objective of our research was to assess to what extent SA methods could be used to identify the letters that belong to companies that were involved in financial misconduct or fraud because the literature has extensively documented that deceptive texts tend to be less positive than truthful ones. The literature on deception notes that such expressions as 'this is not easy' or 'this is difficult', which might be considered equivalent from the semantic point of view, are, in fact, a linguistic unconscious choice that eventually affects the general polarity of the text [2].

However, the literature on linguistic features of deception has mostly described linguistic features of English texts, while the linguistic expression of deception could be, to some extent, language dependent. For example, although some studies have found that the use of passive verbs is more frequent in English deceptive texts, Spanish is a language with particular pragmatic constraints on the use of passive constructions, which are much less common [3]. Therefore, our research first validated that linguistic features of deception most commonly found in English were also characteristics of fraudster's texts in Spanish. The analysis confirmed that Spanish fraudster's texts are also less positive than the other ones, and therefore supported our SA-based approach.

Moreover, note that, usually, SA is applied to social media texts, which are mostly spontaneous texts. Similarly, most of the research on deceptive texts has analyzed emails or phone calls that are personal texts. However, CEOs letters in financial reports are supposed to be prepared over a long period of time and are most likely to be authored in collaboration with a trained writing staff [4]. It had to be proved that the general tone, which is a cue for deception, was kept. Therefore, identifying fraudster's texts in Spanish with SA is a challenging area that, to our knowledge, has not been addressed yet.

The paper is organized as follows. We first review related work on linguistic cues for deception in Section 2. The following sections are devoted to the description of the data sets and how texts were processed. In Section 3.3, we present our analysis to validate that fraudster's texts exhibit the characteristics of deceptive texts. In Section 3.4, we report two SA classification experiments for identifying fraudster's texts. In the first experiment, reported in Section 3.4.1, we compute a general sentiment score to identify fraudster's texts, following the rationale of the Sentiment Orientation Calculator (SO-CAL, [5]). In the second experiment, reported in Section 3.4.2, we built a supervised machine learning classifier to identify the letters of the fraudster companies. The results of both experiments are presented in Section 4 and discussed in Section 5. Finally, the conclusions are presented in Section 6.

2. Related Work

The central presupposition of research on deception detection is that the behavior of people telling lies differs from that of people telling the truth. For example, abnormal physical behavior, such as avoiding eye contact or prolonged hesitations, may indicate that someone is lying [2,6]. In much the same way, differences in text composition and, in particular, in text polarity, were found to be a linguistic indicator of deceptive texts. However, no references to using SA methods for the task of identifying deceptive texts have been found, although SA is also applied to a variety of tasks and languages, with some that are closer to our research, for example, predicting stock market prices [7], or analyzing financial news [8].

Our initial point was reviewing the descriptive research on linguistic cues for identifying deceptive texts to validate that fraudster's letters are deceptive texts. In the literature, the consensus is that English deceptive texts exhibit some common linguistic characteristics. Deceptive texts are longer than those that are non-deceptive, showing less lexical diversity (i.e., fewer repeated words) and lower syntactic complexity [9,10]. Mobility verbs, such as 'go', 'walk' or 'move', were also found to be more frequent in deceptive texts, according to [2,11,12]. It is also a common observation reported by many different authors [2,9,11,13–17], that, in English, liars tend to use fewer self-references (use of first person pronouns or verbal forms) and more negative emotional words.

Research carried out for languages other than English confirmed that many of the linguistic findings just mentioned also hold for such languages as Chinese, Dutch, Italian, German and Spanish (as reported in [18–23], respectively). However, there are some differences as well. We have already mentioned the case of passive sentences in Spanish. Fornaciari and Poesio [20] also found that in Italian, deceptive texts contain more self-references, contrary to English and Dutch texts.

Almela et al. [21] conducted related research that works on Spanish deceptive texts. These authors carried out an experiment with ad hoc written texts in Spanish by participants that were asked to express opinions, including deceptive ones, about a number of topics. The analysis was based on using the LWIC 2001 tool [24]. With this tool, hand-made dictionaries were used to identify and quantify a number of linguistic cues with which to build a classifier (for Spanish resources, see [25]). They reported 70% success in the classification task. The results of Almela et al. showed that the features that performed best for deception identification in Spanish were those that in the LWIC framework are referred to as 'psychological processes'. The features in this category are mostly adjectives that are

an expression of positive or negative emotions: 'feliz' (happy), 'bonito' (pretty), 'hermoso' (beautiful), etc.

There are two works closer to our task of identifying deception in the financial domain, although they are in English texts, again. They demonstrated that emotional words and, in particular, adjectives were found to be good indicators for the task. On the one hand, in [6], Larcker and Zakolyukina used the LWIC tool [24] for developing custom dictionaries to identify and quantify different sentiment expressions. Larcker and Zakolyukina found that, in quarterly conference calls, CEOs of fraudster companies used more extreme positive emotion words, such as 'fantastic' and 'great', and significantly fewer non-extreme positive words, such as 'nice' and 'accept', than non-deceiver CEOs. On the other hand, in [17] Goel and Uzuner investigated the distribution of linguistic features related to the polarity of the text found in particular sections of financial annual reports. Their work also found that such features as the number of positive and negative words and the use of adverbial intensifiers are features related to deception that could also be indicators of fraud. In an experiment similar to ours, they built a number of SVM classifiers, using Weka [26] and different combinations of features. Their corpus contained 180 fraudulent and 180 truthful management and analysis sections of financial annual reports; in addition to lexical features (i.e., words), they used different PoS counts (number of nouns, adjectives, superlative adverbs, etc.) as quantitative features of the different texts to create vector text representations. In their experiments, the classification accuracy ranged from 68.5% when only using lexical information to 73.9% with only data about the frequency of different PoS. When taking the 10 features that performed best from all the available information, accuracy increased to 81.8%. A total of 6 of the top 10 most useful features were frequency counts of different PoS, while the remaining 4 were related to lexical choices.

3. Materials and Methods

3.1. Data and Sample Selection

For our study, a corpus of annual financial reports from different companies was compiled. The publication and wide dissemination of annual financial reports is a legal requirement in Spain in order to promote transparency. The annual financial reports are formal documents with detailed information addressed to shareholders about the financial activities of companies and are meant to be a justification of the management. The information is presented in figures with extensive explanations and a section devoted to management discussion and analysis. Traditionally, they also include, as a foreground, a letter addressed to shareholders, signed by the president or the CEO of the company. We collected these letters, extracting them from publicly available annual reports. We used Spanish reference newspapers (for instance, El País, El Periódico, and El Mundo) and court decisions to identify companies that were sanctioned for accounting fraud or for misrepresentation in financial information during the period from 2011 to 2018. A conservative method was applied in the selection process since only years in which the fraud was proved in a court judgment or publicly accepted were added. Cases in which there were only suspicions were not added. Additionally, we have anonymized the texts to be able to share the data with other researchers. Not all the companies that were found to have committed fraud are represented in our data set because we could not find a digital version of the official annual report, and for some, even the web page has been removed from the world wide web.

The final corpus contains texts from 17 fraudulent and 13 non-fraudulent companies. We manually extracted the CEO's letter addressed to shareholders from the different annual reports selected for the experiments. The corpus size and text length characteristics are described in Table 1.

In general, it was difficult to find fraudster's texts; therefore, the corpus is unbalanced for the non-fraudster's texts and in the experiments, we balanced the corpus to obtain a baseline.

Table 1. Figures describing the corpus used for our experiments in the number of documents, the number of tokens (words) and the document length average in tokens (words).

Corpus	Documents	Tokens	Length Average
Fraudster	35	34,240	951
Non-fraudster	60	85,645	1427
Total	95	119,885	

For the first experiment, as described in Section 3.4.1, 24 fraudster and 24 non-fraudster documents were randomly selected. Additionally, 8 non-fraudster and 7 fraudster documents were selected as held-out data set for testing purposes.

For the second experiment, as described in Section 3.4.2, due to the limited size of the corpus, we used all available documents and we applied class-balancer methods to reweight the instances in the data.

3.2. Text Processing

Annual reports were gathered in PDF format and converted to plain text, encoding with character set UTF-8. Texts were further processed with FreeLing v4.0 (http://nlp.lsi.upc.edu/freeling/ accessed on 27 July 2021) [27] to get the texts tokenized, lemmatized and annotated with Part-of-Speech. The tool Contawords (http://contawords.iula.upf.edu accessed on 27 July 2021) was used to obtain frequency measures of tokens in each document.

3.3. Linguistic Cues

First, the CEO's letters were analyzed to validate whether they exhibit the same linguistic features that the literature reported as characteristic of deceptive texts: (1) length of the texts, (2) lexical diversity or TTR, (3) more third person usage and fewer self-references and (4) use of emotional words, particularly adjectives.

First, deceptive texts were found to be longer than truthful ones [9,10]. In Table 1, we already described the characteristics of the corpus in terms of the document length average. On average, fraudster's texts are shorter (951 tokens per document) than non-fraudster's texts (1427 tokens per document). However, note that 7 documents out of the 35 fraudster documents are longer than the non-fraudster average, and that 18 out of the 60 non-fraudster documents are shorter than the non-fraudster length average.

Second, deceptive texts were found to have lower lexical diversity or richness than truthful ones [9,10]. The Type-Token Ratio (TTR) is the usual measure to assess lexical diversity: the closer the ratio is to 1, the greater the diversity. For our documents, TTR is 0.67 for fraudster's texts and 0.74 for non-fraudster's texts. TTR shows that fraudster's texts in Spanish also tend to have lower lexical diversity, as suggested by the literature for English. However, take into account that a *t*-test significance assessment shows that the difference is not significant (Student *t*-test).

Third, different studies have demonstrated that deceptive texts show fewer self-references, for instance, first person pronouns in English (see [6,12]). For Spanish, which is a non-obligatory subject language, third versus first person reference choices can be better observed in verbal morphology than in pronouns, for instance, the forms 'quiero' (I want), or 'tengo' (I have to), which are normally used for expressing gratitude in such expressions as 'quiero agradecer' (I want to thank) at the beginning or the end of the letters.

We used the PoS annotated version of the corpus to extract the number of verbal forms in third and first, singular and plural. The figures are shown in Table 2 in terms of absolute and relative frequencies.

Table 2. The figures refer to absolute and relative frequencies (RF) of verbal forms of third and first, singular and plural.

Tag	Non-Fraud Doc.	RF	Fraud Doc.	RF
V...3S	688	0.00803	724	0.02114
V...3P	310	0.00361	342	0.00998
V...1S	75	0.00087	49	0.00143
V...1P	239	0.00279	225	0.00657

Relative frequencies show that fraudster Spanish letters do not contain fewer occurrences of first person verb forms as claimed in the deceptive literature. However, note that the figures in Table 2 are quite similar for both classes, and the differences were not found to be statistically significant either.

Finally, many authors have reported that deceptive texts have more emotional words [17] and, in particular, more negative words than non-deceptive texts [2,9–11,13,14,16]. The literature shows that most of the differences in the number of positive and negative words come from the number of adjectives. Accordingly, for assessing the differences in the number of positive and negative words, if any, we counted the occurrence of previously classified positive and negative adjectives by using the Spanish dictionaries provided by the SO-CAL resources [5]. We relied on the SO-CAL resources because they have one of the very few large polarity dictionaries available for Spanish. The SO-CAL Spanish adjective dictionaries that we used for the experiments contain inflected forms or tokens in four different lists that correspond to different sources, including machine-translated English sentiment dictionaries. For our study, we worked with 2200 adjectives—1105 negative and 1095 positive—coming from the SO-CAL non-automatically translated source files to confirm whether the finding in the literature about deceptive texts having more emotional adjectives in general and having fewer positive and more negative ones in particular, also holds for fraudster's texts.

The SO-CAL resources demonstrated high levels of accuracy in classifying the sentiment of texts from a range of domains: news articles, social media comments and blog posts. However, Loughran and McDonald [28] found that almost three-fourths of the negative words in an English general domain word list were not negative in a financial context. For instance, they found that such nouns as 'tax' and 'cost', which are neutral terms in the financial context, were classified as negative in a general domain word list. Moreover, in [28], it was demonstrated that, because of a limited lexical variation (TTR) in these types of texts, very few words account for a large percentage of the total number of polar terms. Thus, in order to prevent the errors pointed out in [28], we decided to work only with adjectives that were found to be less subject to polarity changes depending on the domain [29] and to validate the polarity of the SO-CAL adjective lists, using the method proposed by Hatzivassiloglou et al. [30]. This method is based on the idea that coordination with conjunctions, such as 'and' and 'but', impose a constraint on the semantic orientation of the words they are coordinating. According to this linguistic constraint, adjectives that appear to be linked by an adjoining conjunction, such as 'and' in English or 'y' ('e' as a graphical variant) in Spanish, have the same semantic orientation. On the other hand, adjectives that are coordinated by a disjoining conjunction, such as 'but' in English or 'pero' in Spanish, have opposite orientations. For example, while the combination 'justa y solidaria' (fair and supportive) is found in our corpus and sounds natural, the combination of 'justa pero solidaria' (fair but supportive) would be odd because both being positives cannot be in a 'but' coordination.

In the list of SO-CAL adjectives, we found only four examples for which a change of polarity was necessary. For instance, although 'político' (political) was assigned a negative orientation score in the SO-CAL dictionary, in our texts, it appeared in coordination with exclusively positive adjectives: 'civil' (civil), 'económico' (economic), and 'social' (social).

The other three adjectives that resulted in an opposite polarity after our revision were 'cambiante' (changeable), 'comercial' (marketing, business), and 'simple' (simple).

As shown in Table 3, in our corpus, fraudster's documents do have a higher number of adjectives in total as well as when counting positive and negative ones separately, which confirms the findings of Goel and Uzuner [17] also for financial texts in Spanish.

Table 3. Number of occurrences of positive and negative adjectives in fraudster corpus and the relative frequency (RF).

Adjs. Polarity	Non-Fraud Doc.	RF	Fraud Doc.	RF
Positive	1043	0.01217	1,212	0.03539
Negative	155	0.00180	183	0.00534
Total	1198	0.01398	1,395	0.04074

Indeed, figures in Table 3 show that in our CEO letter corpus, there are more negative adjectives in the fraudster group but also more positive ones, which contradicts the generalized finding for English deceptive texts that claimed that deceptive texts contain fewer positive words and more negative words [2,9–11,13,14,16]. Note that relative frequency (RF) in Table 3 normalizes the number of occurrences because of the different sizes of the compared corpora as shown in Table 1.

3.4. Text Classification

In the previous section, we have seen that none of the linguistic cues that were identified in the literature can be used as a strong predictor of fraudster's texts. The evidence found shows that only a higher number of both positive and negative adjectives seems to support the hypothesis of an objective difference separating fraudster's texts from non-fraudster ones. Therefore, we only used adjectives to assess the extent to which counting their occurrences could achieve a similar accuracy to known attempts to build classifiers as [21] for Spanish deceptive texts and [6] for English sections of fraudster financial reports. We followed the same two mainstream approaches: lexical analysis with the scores of the SO-CAL resources and a bag-of-words representation of texts that records the occurrences of the selected adjectives for training an SVM classifier. The experiments and the results are described in the following sections.

3.4.1. Lexical-Based Sentiment Analysis

The lexical-based method was found to be a good baseline for our machine learning classification experiment because it proved to be effective in analyzing the general tone of texts. The simplicity and straightforwardness of the word list method increases the transparency of the results [28] and, moreover, this method can be said to mirror previous deception studies that looked at frequency counts of positive and negative words like Ref. [21].

In the lexical analysis approach, each word in a dictionary is annotated with a score that ranges from -5 to 5. The SO-CAL method [5] is basically an assessment of the final value after summing these scores. Thus, the final score considers both the orientation and degree of the adjectives in the document. The method described in [5] is more complex and takes into account other linguistic evidence, such as negation and intensifiers, but we have restricted our experiment to the assessment of the adjective values as a first approach. The sentiment score was obtained by first summing the positive and negative adjective scores of each of the 48 randomly chosen documents. The document score was then used for computing the average value for each class. The computed average value was 78.83 for fraud documents and for non-fraud documents, 94.62. This class average score confirmed that fraudster documents, such as deceptive ones, have a lower positive tone than truthful ones.

For creating a classifier with this information, we computed a threshold to separate the documents according to the classes: fraudster and non-fraudster. Thus, for our purpose of using the sentiment score to predict fraudster-related authors, the score average assessed with all our training documents was used as a classification threshold. According to the evidence about fraudster's texts being less positive, the classification rule was (1):

(1) if the sentiment score is lower than average, then the document is a fraudster case; otherwise, it is not.

For refining the results of the classification rule (1), we studied to combine the score with other cues, such as lexical diversity (or Type-Token Ratio) and text length, but with no success in improving the classification results.

For evaluating the lexical-analysis method, we reserved, as explained in Section 3.1, 7 fraudster and 8 non-fraudster documents as the held-out corpus.

3.4.2. Machine Learning

Despite the very successful use of deep-learning methods in the last years, the choice of an SVM machine learning method for our experiment was motivated by the size of our data set, which, as mentioned, is very small. Deep learning methods require very large quantities of data. SVM engines require some research feature selection, such as that which we used for identifying linguistic cues. Recall that Almela et al. [21] and Goel et al. [17] also used a SVM for building a supervised classifier. In our experiment, we represented each document as a bag-of-words of the 633 most frequent adjectives extracted from a corpus made of all sections of the financial reports. This corpus amounted to 482.000 tokens. Note that only 169 adjectives from the final list occur more than 10 times in our texts. As we will see in the discussion, it was important to have document representations that are not too sparse. That is, there can be words that are very important for the class, but occur in very few documents. These cases are not useful for classification [31].

Thus, texts were represented as vectors of 633 components with which we trained an SVM classifier in the Weka [26] implementation of the SMO classifier with the default settings. We used all 95 texts available in a 10-fold cross validation experiment. Later, for comparing the results of the SVM classifier with the ones obtained with the Lexical Classifier, we ran the experiment using a partition of 80% for training and 20% for testing. Additionally, we also applied the Class-Balancer available in Weka to re-weight the instances in the data so that each class had the same total weight.

4. Results

We present now the results of the two experiments. For the sake of comparison, we provide the accuracy and confusion matrices as obtained with the held-out test set used primarily for the lexical-based experiment in Table 4 and with a 80–20% split of the corpus for the SVM classifier in Table 5. For the SVM experiment, results of the cross validation are provided as well in Table 5.

Even with this small training data set, the machine learning based experiment obtained better results than the lexical analysis baseline. The SVM delivered substantially better accuracy. The lexical classifier precision for the fraudster class is 0.66, while the recall is 0.85. In contrast, the SVM classifier precision for the fraudster class is 0.87 and recall 0.77.

As mentioned before, we further tested the SVM classifier with a 10-fold cross-validation evaluation, which delivered lower results—81% accuracy instead of 84%—but still better than the accuracy achieved by the lexical analysis tool.

Table 4. Lexical-analysis classifier results: confusion matrix and accuracy with the held-out test set.

Gold-Standard	Non-Fraud Doc.	Fraudster Doc.
Non-fraud	5	3
Fraudster	1	6
Accuracy	73 %	

Table 5. Results of the SVM Classifier 80%–20% training-test scenario. Confusion matrix and accuracy.

Gold-Standard	Non-Fraud Doc.	Fraudster Doc.
Non-fraud	9	1
Fraudster	2	7
Accuracy	84%	
Balanced-dataset accuracy	83.6%	
Accuracy in 10-fold cross-validation	81%	

5. Discussion

The results of our classification experiments with fraudsters texts are in line with other similar experiments, although they cannot be formally compared with them because they use different data sets. For English texts extracted from financial reports, Goel and Uzuner [17] reported a 81% accuracy in detecting fraudster's texts. In their experiment, described in Section 1, they also used an SVM engine but with very different document representation. They used frequency counts of the different PoS, including number of occurrences of comparative and superlative adverbs. Instead, for representing documents, we used a bag-of-words made of a number of frequent adjectives. With our experiment, we demonstrated that the occurrence of frequent adjectives can be used to classify fraudster's texts with similar results. Moreover, note that the data set of Goel and Uzuner [17] is made of 180 fraudulent and 180 truthful documents, with a 12,843 word token length average per document, while we only had 35 fraudster documents with a length average of 951 word tokens. The letters of our corpus are shorter than their sections of the financial reports, and in the future, we plan to combine letters with other sections of the financial report to improve the results.

Our results can also be related to the research on deception texts in Spanish reported in [21]. Almela et al. also used a bag-of-words representation to build an SVM classifier to identify Spanish deceptive texts. Their research proved that the lexical-analysis classifier built with the LWIC tool achieved better classification than an SVM with which 0.64 F1 for classifying deceptive texts was reported. Our SVM classifier for fraud texts achieved instead 0.82 F1. However, the results of both experiments are not formally comparable: the texts were of a very different genre, as they were written by 100 participants that were asked to express two statements of five sentences minimum on four particular topics expressing truth and deceptive opinions, respectively. Our informal interpretation of the difference in the results is that it could be due to the fact that we used a reduced bag-of-words with only 633 of the most frequent adjectives. It was shown that using a large bag-of-words with short texts (usually the whole vocabulary, as in [21] experiment) often results in data sparsity and poor classification results [32]. For instance, the SVM classifier with the 2200 adjective lists used for the lexical-analysis classifier delivered a 74.7 accuracy.

As for the explainability of the results, the weights assessed by the SVM can be inspected and traced in the different documents as evidence for motivating the classification. The weights assessed for each feature can be interpreted in terms of their importance for drawing the line that separates the classes that the classifier is trying to learn. In Table 6, we show 9 of the 633 features that obtained the most positive weights, and 9 that with the least weight (negative weights). The distribution of these adjectives in the documents of both classes is also shown. From the examples in Table 6, we see that features in the positive top often are only found in non-fraudster's texts, or much more frequently in this class. On the other end, with the lowest weights, there are less important differences for other adjectives that occur in both classes.

The error analysis showed some more cues about the classifier decisions. In the false negative case, the document contained 20 of the 633 devised features. 'Nuevo' (new) is the only one that occurs twice, the other 19 occur only once. Most of these 19 adjectives are in the top 50 features that are important for separating the two classes and which are more frequent in our negative class, i.e., truthful documents. In the case of the false positives, we

observed that they contain between 100 and 200 adjectives each. Recall that the occurrence of more adjectives was the most evident characteristic of fraudster's texts as described in Section 3.3.

Table 6. Sample of more/less weighted adjectives and occurrences in texts.

Features	Non-Fraud Doc.	Fraudster Doc.
extraordinarios (extraordinary)	3	0
cordial (cordial)	3	0
corporativos (corporate)	3	0
perfecta (perfect)	1	0
alemana (German)	4	0
favorables (favorable)	11	0
optimistas (optimistic)	11	0
industriales (industrial)	11	1
internacionales (international)	39	16
comercial (commercial)	21	6
productivas (productive)	0	2
estratégica (strategic)	6	2
valiosa (valuable)	2	1
español (Spanish)	9	7
complejas (complex)	0	1
principal (main)	13	17
constante (constant)	9	10
prometedor (promising)	2	2

6. Conclusions

The aim of our study was to assess to what extent sentiment-related information could be used to identify letters that are from companies that have been involved in financial misconduct or fraud. The evidence found showed that only a higher number of both positive and negative adjectives support the hypothesis of an objective difference in the fraudster's texts. Our experiments also demonstrated that using a document representation based on a bag-of-words with about 600 frequent adjectives captures the frequency and combination of adjectives, which are operative cues for identifying letters from companies that have committed fraud with more than 83% accuracy. Our experiments also show that the decision made by the automatic classifier can be inspected and explained in relation to significant characteristics that characterize the classes: fraudster's documents tend to contain more adjectives than non-fraudster's ones. However, the motivation under the particular use of particular adjectives or their combination must be sought in psychological studies.

Author Contributions: Data curation, G.B. and S.A.; investigation, N.B., G.B. and S.A.; methodology, N.B.; resources, S.A.; software, N.B.; writing—original draft, N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Spanish Plan Estatal de Investigación Científica y Técnica y de Innovación 2017–2020, PID2019-104512GB-I00.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The corpus will be made available at the institutional UPF electronic repository. For the publication time a URL will be provided. Data will be shared only for research purposes and with a license signature to prevent dissemination for unwanted purposes.

Acknowledgments: We are sincerely grateful to Oriol Amat who suggested the topic of the research, where to find data and made other very important suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cambria, E. Affective Computing and Sentiment Analysis. *IEEE Intell. Syst.* **2016**, *31*, 102–107. [CrossRef]
2. Newman, M.L.; Pennebaker, J.W.; Berry, D.S.; Richards, J.M. Lying Words: Predicting Deception from Linguistic Styles. *Personal. Soc. Psychol. Bull.* **2003**, *29*, 665–675. [CrossRef] [PubMed]
3. Quesada, J.D. Obituary: Adios to passive in Spanish. *La linguistique* **1997**, *33*, 41–62.
4. Humpherys, S.L.; Moffitt, K.C.; Burns, M.B.; Burgoon, J.K.; Felix, W.F. Identification of fraudulent financial statements using linguistic credibility analysis. *Decis. Support Syst.* **2011**, *50*, 585–594. [CrossRef]
5. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-Based Methods for Sentiment Analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [CrossRef]
6. Larcker, D.F.; Zakolyukina, A.A. Detecting deceptive discussions in conference calls. *J. Account. Res.* **2012**, *50*, 495–540. [CrossRef]
7. Gupta, R.; Chen, M. Sentiment Analysis for Stock Price Prediction. In Proceedings of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, China, 6–8 August 2020; pp. 213–218. [CrossRef]
8. Štrimaitis, R.; Stefanovič, P.; Ramanauskaitė, S.; Slotkienė, A. Financial Context News Sentiment Analysis for the Lithuanian Language. *Appl. Sci.* **2021**, *11*, 4443. [CrossRef]
9. Burgoon, J.K.; Buller, D.B.; Floyd, K.; Grandpre, J. Deceptive realities: Sender, receiver, and observer perspectives in deceptive conversations. *Commun. Res.* **1996**, *23*, 724–748. [CrossRef]
10. Burgoon, J.; Stoner, G.; Bonito, J.; Dunbar, N. Trust and deception in mediated communication. In Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 6–9 January 2003. [CrossRef]
11. Zhou, L.; Twitchell, P.L.; Qin, T.; Burgoon, J.K.; Nunamaker, J.F. An exploratory study into deception detection in text-based computer mediated communication. In Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 6–9 January 2003.
12. Mihalcea, R.; Strapparava, C. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In Proceedings of the ACL-IJCNLP 2-7, Singapore, 4 August 2009; pp. 309–312.
13. Goel, S.; Gangolly, J. Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Intell. Syst. Account. Financ. Manag.* **2012**, *19*, 75–89. [CrossRef]
14. Hancock, J.T.; Curry, L.E.; Goorha, S.; Woodworth, M. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Process.* **2007**, *45*, 1–23. [CrossRef]
15. Hobson, J.L.; Mayew, W.J.; Venkatachalam, M. Analyzing Speech to Detect Financial Misreporting. *J. Account. Res.* **2012**, *50*, 349–342. [CrossRef]
16. Liu, X.; Hancock, J.; Zhang, G.; Xu, R.; Markowitz, D.; Bazarova, N. Exploring linguistic features for deception detection in unstructured text. In Proceedings of the 45th Hawaii International Conference on System Sciences, Hawaii, HI, USA, 4–7 January 2012.
17. Goel, S.; Uzuner, O. Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intell. Syst. Account. Financ. Manag.* **2016**, *23*, 215–239. [CrossRef]
18. Zhou, L.; Sung, Y.W. Cues to deception in online Chinese groups. In Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), Waikoloa, Hawaii, 7–10 January 2008. [CrossRef]
19. Schelleman-Offermans, K.; Merckelbach, H. Fantasy proneness as a confounder of verbal lie detection tools. *J. Investig. Psychol. Offender Profiling* **2010**, *7*, 247–260. [CrossRef]
20. Fornaciari, T.; Poesio, M. Automatic deception detection in Italian court cases. *Artif. Intell. Law* **2013**, *21*, 303–340. [CrossRef]
21. Almela, A.; Valencia-garcía, R.; Cantos, P. Seeing through deception: A computational approach to deceit detection in written communication. *Linguist. Evid. Secur. Law Intell.* **2012**, *1*, 15–22. [CrossRef]
22. Hauch, V.; Blandón-Gitlin, I.; Masip, J.; Sporer, S.L. Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personal. Soc. Psychol. Rev.* **2015**, *19*, 307–342. [CrossRef] [PubMed]
23. Masip, J.; Bethencourt, M.; Lucas, G.; Segundo, M.S.S.; Herrero, C. Deception detection from written accounts. *Scand. J. Psychol.* **2012**, *53*, 103–111. [CrossRef] [PubMed]
24. Pennebaker, J.W.; Francis, M.E.; Booth, R.J. *Linguistic Inquiry and Word Count (LIWC)*; Lawrence Erlbaum Publisher: Mahwah, NJ, USA, 2001.
25. Ramírez-Esparza, N.; Pennebaker, J.W.; García, F.; Suriá, R. La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista Mexicana de Psicologia* **2007**, *24*, 85–99.
26. Frank, E.; Hall, M.A.; Witten, I.H. The WEKA Workbench. *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4th ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2016.
27. Padró, L.; Stanilovsky, E. FreeLing 3.0: Towards Wider Multilinguality. In Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, 21–27 May 2012.
28. Loughran, T.; McDonald, B. Textual Analysis in Accounting and Finance: A Survey. *J. Account. Res.* **2016**, *54*, 1187–1230. [CrossRef]
29. Vázquez, S.; Bel, N. A Classification of Adjectives for Polarity Lexicons Enhancement. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 21–27 May 2012.

30. Hatzivassiloglou, V.; McKeown, K.R. Predicting the Semantic Orientation of Adjectives. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 7–12 July 1997.
31. Bel, N. Handling of Missing Values in Lexical Acquisition. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 17–23 May 2010.
32. Colas, F.; Paclík, P.; Kok, J.N.; Brazdil, P. Does SVM Really Scale Up to Large Bag of Words Feature Spaces? In *Advances in Intelligent Data Analysis VII*; Berthold, R.M., Shawe-Taylor, J., Lavrač, N., Eds.; Springer: Berlin, Germany, 2007.