

UNIVERSITY
OF TWENTE.



Marcia Fissette

11-04-2018

Jaarverslag

Verslag van de activiteiten van een bedrijf van het voorgaande jaar.

Verslag bevat

Jaarrekening

- Balans
- Winst- en verliesrekening
- Kasstroomoverzicht

Tekstuele informatie

- Toelichting op de jaarrekening
- Brief aan de aandeelhouders
- Verslag van de accountant
- *Management Analyse & Discussie*

Fraude in jaarverslagen

Financiële effect

Grootste financiële effect van alle fraude types

Voorbeelden:

- Enron: Aandelen gezakt van \$90,56 per aandeel naar centen
- Parmalat: schuld van 14 miljard en 300 miljoen euro (8 keer hoger dan gerapporteerd)

Frauduleuze activiteiten

Vervalsing van onderliggende documenten

Vervalste schattingen en oordelen

Achterhouden van belangrijke informatie

Onderzoeksvraag

Kan een **tekst mining** model worden ontwikkeld die **indicaties van fraude** kan detecteren in **jaarverslagen** van bedrijven **wereldwijd**?

Tekst mining

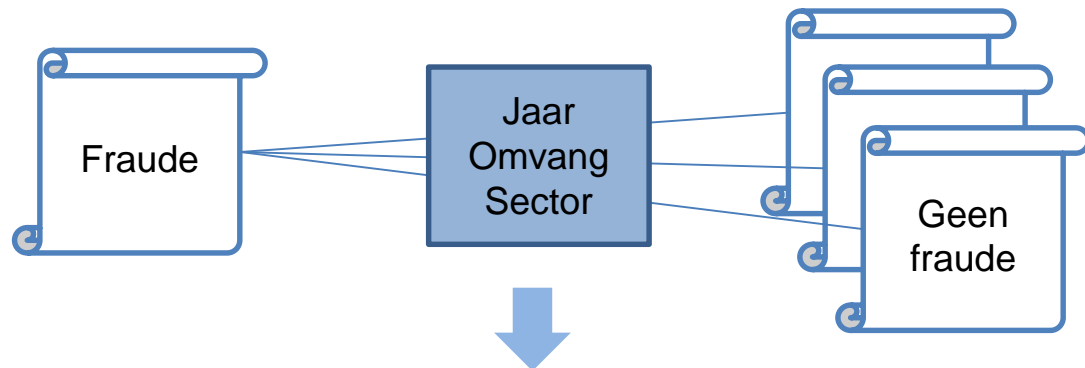
Automatisch identificeren van patronen in tekst

Model bestaat uit

1. Een data set bestaande uit teksten
2. Procedure om tekst te transformeren naar een gestructureerde representatie
3. Machine learning procedure om patronen te leren

Gebruikt om te classificeren

De data



402 frauduleuze jaarverslagen
1.325 niet frauduleuze jaarverslagen

Data voorbereiding

Verwijderen van html tags

Uitsluiten van tabellen

Kleine letters

Tokenization

Zinnen

Woorden

Verwijderen van interpunctie

Features

Unigrams

Bigrams

Beschrijvend

- Tekst lengte
- Lexicale diversiteit

Complexiteit

- Zin complexiteit
- Woord complexiteit

Grammaticaal

- Part-of-speech woord categoriën
- Werkwoordstijd

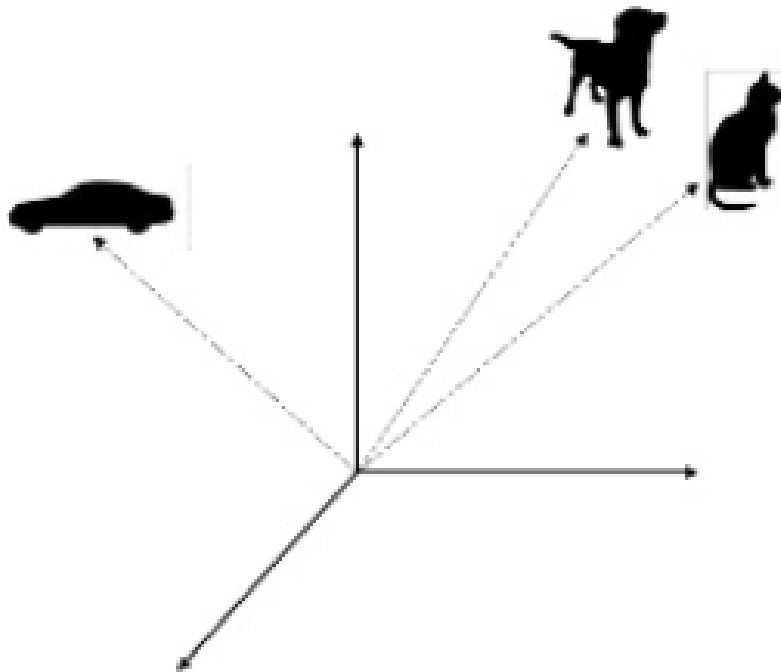
Leesbaarheid

- Jaren onderwijs nodig

Psychologisch

- Positief vs. negatief
- Zekerheid

Woord vectoren



Machine learning

Split data

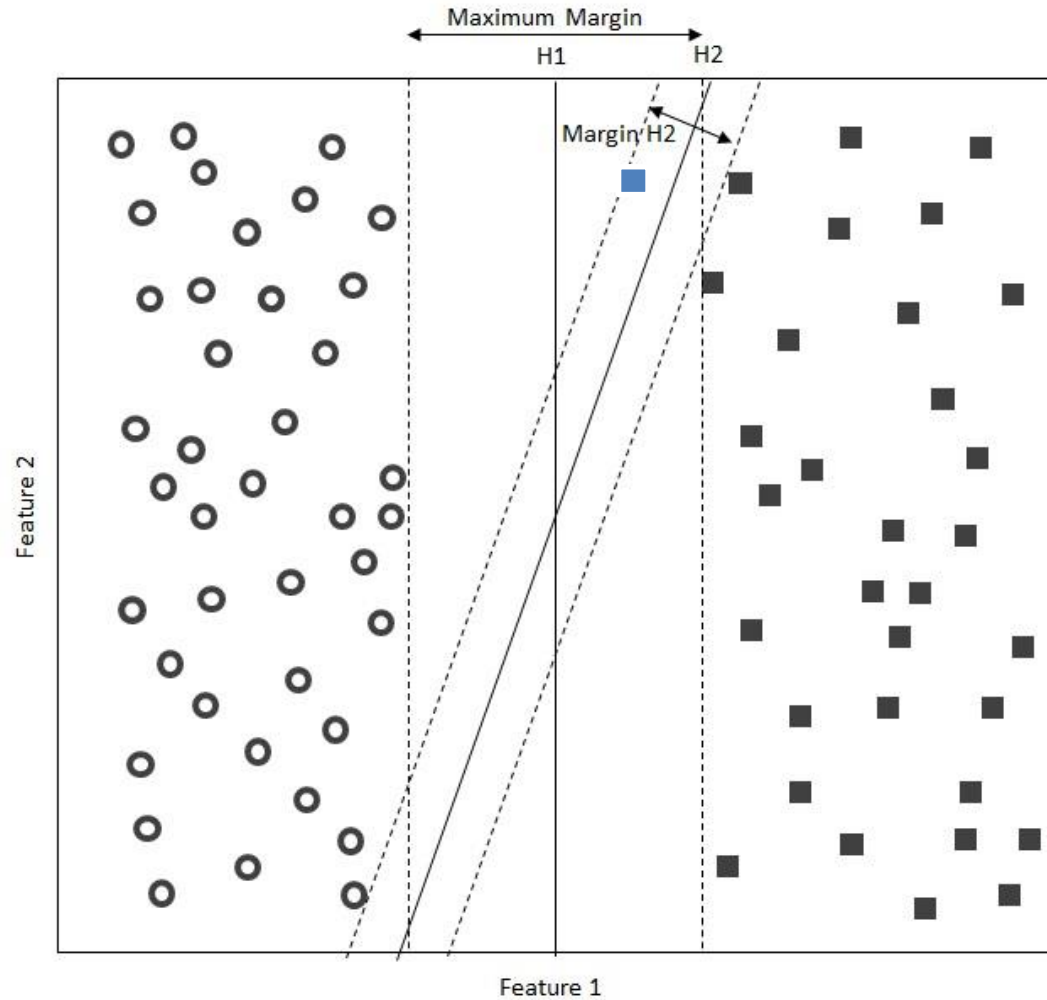
Development– validation: 70% - 30%

Naïve Bayes

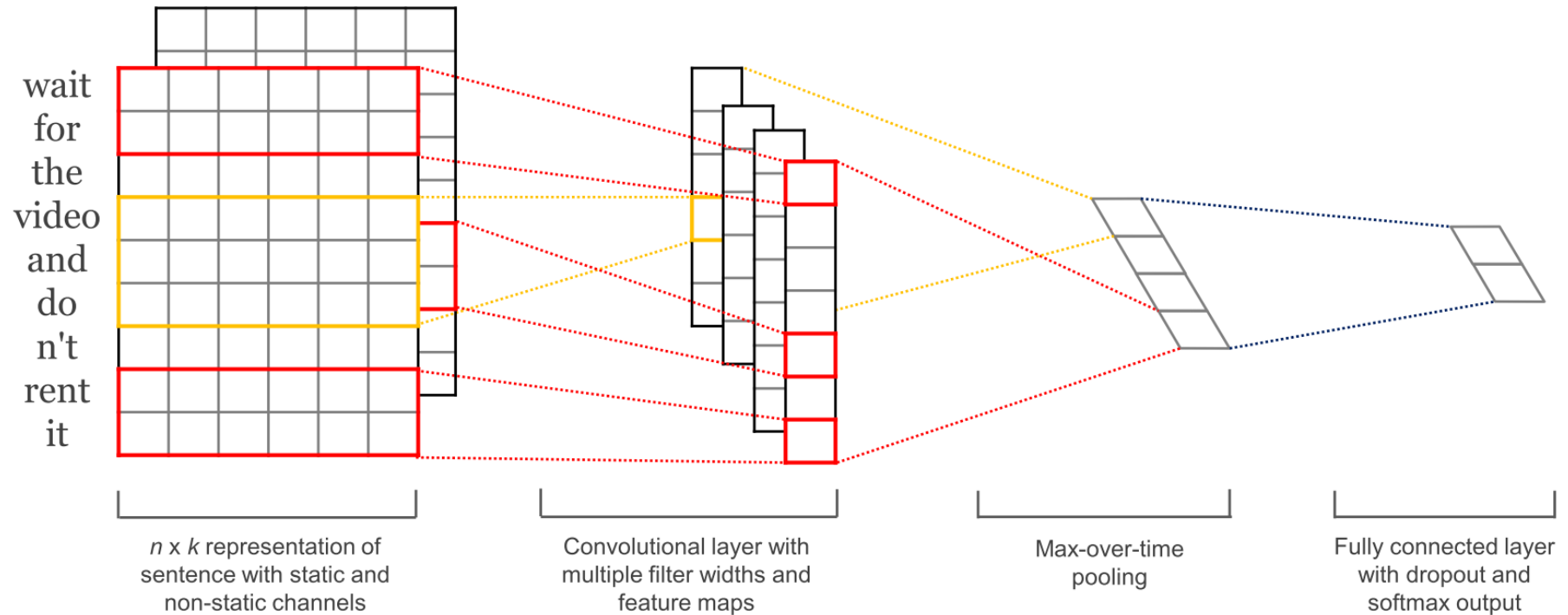
Support Vector Machine

Convolutional Neural Network

Support Vector Machine



Convolutional Neural Network



Resultaat

Kan een **tekst mining** model worden ontwikkeld die **indicaties van fraude** kan detecteren in **jaarverslagen** van bedrijven **wereldwijd**?

Tekst mining kan bijdragen aan fraudedetectie

Naïve Bayes model presteert beter dan Support Vector Machine in fraude herkenning

Unigrams beste features

Convolutional Neural Network potentie om beter te presteren dan Naïve Bayes

fraudulent activities. The research described in this thesis examined the contribution of text analysis to detecting indications of fraud in the annual reports of companies worldwide. In an annual report, a company presents its financial results and activities from the previous year. In addition, the annual report contains textual explanations. In case of fraud, the annual report does not provide a fair view of the financial position of the company. For the research described in this thesis, a total of 1,727 annual reports have been collected, of which 402 are of the years and companies in which fraudulent activities took place, and which have an impact on the information disclosed in the annual report. Since it is assumed that most companies do not engage in such fraudulent activities, the majority, 77%, of the annual reports in the data set are not fraudulent. Furthermore, the composition of this data set takes into account the possibility that the year the annual report is concerned with, the sector in which the company operates and the size of the company may affect the information disclosed in the annual report. Therefore, for each fraud annual report, non fraudulent annual reports from the same year, of a company in the same sector, and of a comparable size are added to the data set. A method for the automatic extraction of information from annual reports has been proposed to obtain the data needed to compile the data set. By applying this method, the year, sector and size of the company can be determined for a large number of annual reports. The approach has also been used to extract the Management Discussion & Analysis (MD&A) section, which is the part of the annual report the research in this thesis focuses on. In the MD&A section, the company provides information concerning the performance and the activities of the preceding year and the expectations for the following year. The first models developed for the research described in this thesis, analyze the texts by counting the words (unigrams) in the MD&A section. These word counts are normalized by the term-inverse document frequency (TF-IDF) method that takes into account the length of the text and the frequency with which a word occurs in the data set. The most informative words are determined using the chi-square feature selection model. This representation of the text is the input of the machine learning algorithms Naive Bayes (NB) and Support Vector Machine (SVM). These algorithms learn patterns to classify the annual reports as 'fraud' or 'no fraud'. The NB model classifies the texts based on probability calculations. The SVM model uses unigrams to create a vector space. Subsequently, the SVM model determines the most optimal hyperplane in the space that separates the 'fraud' and 'no fraud' texts. The NB model shows the best performance. The percentage of correctly classified annual reports is 89%. Subsequently, the NB and SVM models based on unigrams are expanded with the linguistic features of the text found to be informative in the previous research concerning the detection of fraud or deception in text. The linguistic features into six categories. The first category comprises groups of two consecutive words (bigrams). The second category consists of features that describe the general properties of the text, such as the total numbers in the text. The third category describes the complexity of the text using measures for the complexity of the words and sentences. The fourth category focuses on the grammatical aspects of the text. The fifth category consists of measures that determine the readability of a text, expressed as the number of years of education needed to understand the text. The final category concerns the Linguistic Inquiry and Word Count (LIWC) tool that is used to extract psychological features, such as emotions, from the text. Furthermore, we developed a new type of feature that reflects the grammatical relations between words. The classification results show that only the addition of bigrams to the SVM model improves the result slightly. The other categories of linguistic features do not improve the result. The latest development in machine learning is the use of deep learning. By using networks consisting of several layers complex patterns can be found. A Convolutional Neural Network (CNN) model has been found successful in research that classifies texts in domains other than fraud. With this model, each word in the text is represented by a vector (word embedding). Word embeddings aim to include the semantic relationships between words, in addition to the representation of the individual words. Due to the limited computer capacity available during the research described in this thesis, experimentation with the CNN model was performed using 40% of the original data set. In order to compare the results with the previous models, an NB model was also developed on this smaller data set. The results are significantly lower than the NB model, but also show that the CNN model achieves slightly better results than the NB model. The results show that text analysis can contribute to the detection of indications of fraud. However, to further enhance the performance of the models, more research may be required either by providing additional sources of information. Future research may experiment with machine learning algorithms, such as the CNN or a combination of various algorithms. The model for text can be expanded with models that use the company's financial information. More textual information may be added to the model from the annual report itself or from other documents of the company. To catch fraudsters and prevent damage caused by financial fraud, innovative fraud detection methods are required that are capable of identifying indications of the continuously changing fraudulent activities. The research described in this thesis examined the contribution of text analysis to detecting indications of fraud in the annual reports of companies worldwide. In an annual report, a company presents its financial results and activities from the previous year. In addition, the annual report contains textual explanations. In case of fraud, the annual report does not provide a fair view of the financial position of the company. For the research described in this thesis, a total of 1,727 annual reports have been collected, of which 402 are of the years and companies in which fraudulent activities took place, and which have an impact on the information disclosed in the annual report. Since it is assumed that most companies do not engage in such fraudulent activities, the majority, 77%, of the annual reports in the data set are not fraudulent. Furthermore, the composition of this data set takes into account the possibility that the year the annual report is concerned with, the sector in which the company operates and the size of the company may affect the information disclosed in the annual report. Therefore, for each fraud annual report, non fraudulent annual reports from the same year, of a company in the same sector, and of a comparable size are added to the data set. A method for the automatic extraction of information from annual reports has been proposed to obtain the data needed to compile the data set. By applying this method, the year, sector and size of the company can be determined for a large number of annual reports. The approach has also been used to extract the Management Discussion & Analysis (MD&A) section, which is the part of the annual report the research in this thesis focuses on. In the MD&A section, the company provides information concerning the performance and the activities of the preceding year and the expectations for the following year. The first models developed for the research described in this thesis, analyze the texts by counting the words (unigrams) in the MD&A section. These word counts are normalized by the term-inverse document frequency (TF-IDF) method that takes into account the length of the text and the frequency with which a word occurs in the data set. The most informative words are determined using the chi-square feature selection model. This representation of the text is the input of the machine learning algorithms Naive Bayes (NB) and Support Vector Machine (SVM). These algorithms learn patterns to classify the annual reports as 'fraud' or 'no fraud'. The NB model classifies the texts based on probability calculations. The SVM model uses unigrams to create a vector space. Subsequently, the SVM model determines the most optimal hyperplane in the space that separates the 'fraud' and 'no fraud' texts. The NB model shows the best performance. The percentage of correctly classified annual reports is 89%. Subsequently, the NB and SVM models based on unigrams are expanded with the linguistic features of the text found to be informative in the previous research concerning the detection of fraud or deception in text. The linguistic features into six categories. The first category comprises groups of two consecutive words (bigrams). The second category consists of features that describe the general properties of the text, such as the total numbers in the text. The third category describes the complexity of the text using measures for the complexity of the words and sentences. The fourth category focuses on the grammatical aspects of the text. The fifth category consists of measures that determine the readability of a text, expressed as the number of years of education needed to understand the text. The final category concerns the Linguistic Inquiry and Word Count (LIWC) tool that is used to extract psychological features, such as emotions, from the text. Furthermore, we developed a new type of feature that reflects the grammatical relations between words. The classification results show that only the addition of bigrams to the SVM model improves the result slightly. The other categories of linguistic features do not improve the result. The latest development in machine learning is the use of deep learning. By using networks consisting of several layers complex patterns can be found. A Convolutional Neural Network (CNN) model has been found successful in research that classifies texts in domains other than fraud. With this model, each word in the text is represented by a vector (word embedding). Word embeddings aim to include the semantic relationships between words, in addition to the representation of the individual words. Due to the limited computer capacity available during the research described in this thesis, experimentation with the CNN model was performed using 40% of the original data set. In order to compare the results with the previous models, an NB model was also developed on this smaller data set. The results are significantly lower than the NB model, but also show that the CNN model achieves slightly better results than the NB model. The results show that text analysis can contribute to the detection of indications of fraud. However, to further enhance the performance of the models, more research may be required either by providing additional sources of information. Future research may experiment with machine learning algorithms, such as the CNN or a combination of various algorithms. The model for text can be expanded with models that use the company's financial information. More textual information may be added to the model from the annual report itself or from other documents of the company. To catch fraudsters and prevent damage caused by financial fraud, innovative fraud detection methods are required that are capable of identifying indications of the continuously changing fraudulent activities.

UNIVERSITY
OF TWENTE.



Marcia Fissette