MAB MAANDBLAD VOOR ACCOUNTANCY EN BEDRIJFSECONOMIE

Maandblad voor Accountancy en Bedrijfseconomie 92(5/6) (2018): 147–155
DOI 10.5117/mab.92.25968

Research Article

# Fraudedetectie door tekstanalyse van jaarverslagen

Marcia Fissette, Bernard Veldkamp, Theo de Vries

**Received** 18 April 2018 | **Accepted** 30 June 2018 | **Published** 23 July 2018

## Samenvatting[1]

Innovatieve fraudedetectiemethoden zijn nodig om aanwijzingen van fraude op tijd te signaleren en verdere schade te voorkomen. De focus van fraudeonderzoek in jaarverslagen is vaak gericht op de numerieke informatie. Jaarverslagen bevatten tekstuele informatie die ook indicaties van fraude bevat. Dit artikel beschrijft het onderzoek naar een methode voor tekstanalyse die gebruik maakt van *machine learning* om jaarverslagen automatisch te classificeren in de twee categorieën 'fraude' en 'geen fraude'.

## Relevantie voor de praktijk

Het effect van fraude in het jaarverslag is groot. Naast enorme financiële schade is er sprake van reputatieschade voor het bedrijf en worden andere partijen zoals de accountants, controllers, advocaten en wederpartijen meegezogen. Om schade in te perken of te voorkomen is het van belang om fraude op tijd te signaleren. Behalve numerieke informatie bevat een jaarverslag tekstuele informatie. Deze teksten alleen al kunnen met grote waarschijnlijkheid de aan- of afwezigheid van fraude in jaarverslagen signaleren. De hier beschreven methode is relevant omdat er geen aparte financiële analyse benodigd is. Pas als er vermoeden van fraude is zal die, indien nodig, moeten worden ingezet.

## 1. Inleiding

Fraude is een wereldwijd fenomeen dat kan voorkomen in alle soorten bedrijven. Zo waren er in het verleden grote schandalen bij energiebedrijf Enron in de VS, zuivel- en voedselproducent Parmalat in Italië en optica- en reprografiebedrijf Olympus in Japan. Hoewel de totale omvang van de kosten van fraude wereldwijd niet met voldoende precisie kan worden berekend, geven schattingen aan dat fraude resulteert in het verlies van 5% van de uitgaven van organisaties (Gee et al. 2017). Wereldwijd betekent dat een omvang van ca. 4,4 biljoen dollar per jaar.

Om financiële fraude op te sporen zijn diverse methoden ontwikkeld. De focus ligt daarbij op kwantitatieve informatie en vooraf gedefinieerde risicofactoren. Een verscheidenheid aan financiële ratio's is ontwikkeld om de financiële positie van een bedrijf te meten (Persons 1995; Spathis et al. 2002; Kaminski et al. 2004; Kirkos et al. 2007). Voorbeelden van risicofactoren zijn het aantal dochterondernemingen van een bedrijf, of het bedrijf gewisseld is van CEO en het hebben van een slechte reputatie (Fanning and Cogger 1998; Bell and Carcello 2000).

De afgelopen jaren is er in het wetenschappelijk onderzoek een verschuiving te zien van de focus op kwantitatieve informatie en risicofactoren naar de tekstuele informatie (Cecchini et al. 2010; Glancy and Yadav 2011; Purda and Skillicorn 2015). Tekst is om verschillende redenen een interessante bron voor onderzoek naar methoden voor fraudedetectie. Ten eerste geeft tekst informatie die aanvullend is op de financiële informatie. In tegenstelling tot de financiële gegevens die informatie geven over het afgelopen jaar, kan de tekstuele informatie ingaan op de verwachtingen en plannen voor de toekomst. Dit is met name het geval in het bestuursverslag. Het toevoegen van de toekomstgerichte informatie aan fraudedetectiemethoden kan mogelijk het resultaat van de methoden verbeteren. Ten tweede bereikt tekst een groter publiek omdat tekst door meer mensen te begrijpen is dan de cijfers in de jaarrekening. De tekstuele informatie heeft daardoor de potentie om meer mensen te beïnvloeden, en in het geval van fraude, te misleiden. Bedrijven zijn zich steeds meer bewust van de invloed die het jaarverslag kan hebben. Waar het jaarverslag eerst vooral gebruikt werd om boekhoudkundige informatie over te brengen, wordt het tegenwoordig steeds meer gebruikt om de bedrijfsidentiteit te presenteren (Beattie et al. 2008; Lee 1994). In de afgelopen decennia is de hoeveelheid tekst toegenomen. De toename van 375% is voornamelijk toe te schrijven

aan de vrijwillige informatieverstrekking (Beattie et al. 2008; Goel and Gangolly 2012).

De toename van tekstuele informatie gaat hand in hand met de toename van computercapaciteit. Deze heeft de mogelijkheden voor geautomatiseerde tekstanalyse sterk verbeterd. Computers kunnen een veel hoger aantal jaarverslagen verwerken in kortere tijd dan mensen. Het handmatig analyseren van teksten is tijdrovend. Computers bieden daarom een uitkomst. Hoewel computers niet zoals mensen de tekst zullen begrijpen, zijn zij wel in staat om de abstracte taalkundige informatie te extraheren. Voor mensen is het juist moeilijk om de feitelijke inhoud van een tekst te negeren en alleen te focussen op hoe iets wordt gezegd (Pennebaker et al. 2003). Dit type informatie kan juist indicaties van fraude bevatten die mensen niet signaleren.

Door de computercapaciteit, het belang van tekstuele informatie en de mogelijke toegevoegde waarde ten opzichte van de financiële gegevens, is het interessant om de mogelijkheden van tekstanalyse voor het detecteren van indicaties van fraude in jaarverslagen van bedrijven zo breed mogelijk, wereldwijd, te onderzoeken.

Paragraaf 2 geeft een overzicht van eerder onderzoek naar fraudedetectie in jaarverslagen en fraude en leugendetectie door middel van tekstanalyse. Paragraaf 3 beschrijft de onderzoeksmethode. De resultaten van het onderzoek worden weergegeven in paragraaf 4. Ten slotte geeft paragraaf 5 de conclusie en een discussie van de resultaten.

# 2. Theoretisch raamwerk

### 2.1 Fraudedetectie door middel van management- en financiële informatie

Modellen voor fraudedetectie zijn veelal gebaseerd op de prikkels en risicofactoren die zijn geïdentificeerd in fraudeonderzoeken. Ndofor et al. (2015) ontdekten dat wanneer het management aandelen bezit, de kans op financiële fraude toeneemt. Het risico op fraude is ook groter wanneer een bedrijf een bestuur heeft dat wordt gedomineerd door een paar individuen (Wang et al. 2011). Agostini and Favero (2013) stellen dat fraude in de jaarrekening het gevolg is van druk die wordt uitgeoefend door investeerders en analisten. Ten slotte laten onderzoeken zien dat de kans op fraude hoger is naarmate de financiele resultaten van bedrijven verslechteren (Beasley et al. 2010; Rezaee 2005).

Onderzoekers hebben de risicofactoren verwerkt in fraudedetectiemodellen. Dergelijke modellen bevatten bijvoorbeeld variabelen vanuit de jaarrekening die de financiële conditie meten (Persons 1995). Beneish (1997) gebruikt daarnaast het percentage aandelen van het management als variabele voor het detecteren van manipulatie van de Generally Accepted Accounting Principles (GAAP). Verschillende onderzoekers ontwikkelden modellen met financiële ratio's, die een verhouding tussen twee waardes uit de jaarrekening weergeven, voor het detecteren van indicaties van fraude in jaarverslagen (Hoogs et al. 2007; Huang et al. 2012; Kaminski et al. 2004; Perols 2011; Ravisankar et al. 2011). Grove and Basilisco (2008) tonen aan dat de financiële informatie alleen niet voldoende is om indicaties van fraude op te sporen. Zij combineren daarom de financiële ratio's met informatie over de corporate governance. De beperkte detectiemogelijkheden van alleen financiële ratio's zouden ook kunnen worden verbeterd door het meenemen van de tekstuele informatie.

### 2.2 Tekstanalyse in financiële verslagen

De hoeveelheid tekst in het jaarverslag is de afgelopen decennia flink toegenomen (Beattie et al. 2008; Goel and Gangolly 2012). Deze toename is voor het grootste deel toe te schrijven aan het vrijwillig vrijgeven van informatie. Als gevolg van deze ontwikkeling hebben onderzoekers de teksten geanalyseerd voor verschillende doeleinden. Li (2006) gebruikt de frequentie van woorden gerelateerd aan risico en onzekerheid in jaarverslagen op form 10-K[2] om lagere toekomstige winsten te voorspellen. Een ander onderzoek ontdekt dat managers de concurrentie-informatie in het bestuursverslag verdraaien (Li et al. 2011). Smith and Taffler (1999) vinden dat negatieve resultaten worden verklaard in technische boekhoudkundige termen terwijl positieve resultaten in termen van oorzaak en gevolg worden uitgelegd. Clatworthy and Jones (2003) constateren dat management de neiging heeft om de externe omgeving de schuld te geven van negatieve resultaten en tegelijkertijd de eer op zich te nemen voor de goede resultaten. Verschillende onderzoekers onderzoeken de mogelijkheid om tekst te gebruiken voor het voorspellen van faillissementen (Balakrishnan et al. 2010; Cecchini et al. 2010; Hájek and Olej 2013; Smith and Taffler 1999).

De toename in de hoeveelheid tekst bevordert de mogelijkheid om tekst te gebruiken in fraudedetectieonderzoek. Voor de ontwikkeling van fraudedetectiemethoden die gebruik maken van tekst kan worden geprofiteerd van de kennis die is opgedaan in onderzoeken naar de detectie van leugens. Net als bij fraude is bij liegen sprake van een opzettelijke poging om anderen te misleiden. Burgoon et al. (2003) analyseren chat-teksten en concluderen dat taalgebruik van leugenaars verschilt van dat van waarheidsvertellers. Verschillende onderzoekers tonen aan dat tekstuele kenmerken geschikt kunnen zijn voor het automatisch detecteren van leugens (Newman et al. 2003; Zhou et al. 2004). DePaulo et al. (2003) ontdekken bijvoorbeeld dat leugenaars minder details geven dan de waarheidsvertellers.

Een mogelijk probleem bij het toepassen van methoden uit het leugendetectieonderzoek op financiële verslagen is dat dit type document door meerdere personen kan zijn geschreven waarbij mogelijk niet iedereen direct betrokken is bij de fraude. De schrijver van de verslagen is misschien niet altijd degene die bewust liegt. Desalniettemin

hebben onderzoekers de technieken voor leugendetectie getest voor de detectie van fraude in 10-K jaarverslagen (Humpherys et al. 2011; Skillicorn and Purda 2012). Wang and Wang (2012) tellen woorden die relevant zijn voor leugendetectie voor het detecteren van fraude in een zeer gelimiteerde dataset met vijf fraudezaken. Goel and Gangolly (2012) vinden dat frauduleuze en niet-frauduleuze 10-K jaarverslagen significant verschillen op complexiteit van zinnen, leesbaarheid en het gebruik van negatieve woorden, passief taalgebruik en woorden die duiden op onzekerheid. Naast het direct overnemen van de kenmerken uit het leugendetectieonderzoek hebben onderzoekers deze kenmerken gecombineerd met patronen die door middel van *machine learning* automatisch uit de tekst worden geëxtraheerd voor het detecteren van fraude in 10-K jaarverslagen of 10-Q kwartaalverslagen (Cecchini et al. 2010; Glancy and Yadav 2011; Goel et al. 2010; Purda and Skillicorn 2010, 2015). Het onderzoek beschreven in dit artikel heeft de meeste raakvlakken met deze laatstgenoemde onderzoeken. In plaats van alleen 10-K jaarverslagen zijn jaarverslagen van bedrijven wereldwijd gebruikt. Daarnaast is een grotere dataset gebruikt. De genoemde onderzoeken gebruiken een dataset waarbij 50% van de verslagen frauduleus is. De dataset is een steekproef van de totale populatie van jaarverslagen. Wanneer de dataset een goede weergave van de werkelijkheid is, is de kans groter dat deze steekproef representatief is voor de gehele populatie jaarverslagen. Dan is het aannemelijk dat de resultaten verkregen met de dataset ook gelden voor de gehele populatie jaarverslagen. Paragraaf 3 beschrijft de volledige onderzoeksmethode.

# 3. Onderzoeksmethode

voor het onderzoek beschreven in dit artikel is een aantal *text mining*-modellen ontwikkeld. Een dergelijk model kan worden opgesplitst in drie delen. Het eerste deel is de dataset bestaande uit tekstuele documenten. De dataset voor dit onderzoek wordt beschreven in paragraaf 3.1. Ten tweede is een procedure nodig voor het omzetten van de tekst naar een gestructureerde representatie die een computer kan verwerken. Dit wordt *feature extraction and selection* genoemd. Paragraaf 3.2 gaat hier kort op in. Ten slotte kan de gestructureerde representatie aan een *machine learning*-algoritme worden gegeven die patronen kan leren. Op basis van de patronen kan een algoritme bepalen tot welke categorie een document behoort. Voor fraudedetectie worden twee categorieën onderscheiden: 'fraude' en 'geen fraude'. In paragraaf 3.3 worden de *machine learning*-algoritmen die gebruikt zijn in dit onderzoek toegelicht.

### 3.1 De dataset

De dataset bestaat uit jaarverslagen van bedrijven wereldwijd, alle Engelstalig, waarvan het merendeel jaarverslagen betreft van bedrijven die op Amerikaanse beurzen genoteerd zijn. Voor de ontwikkeling van een model dat jaarverslagen toekent aan de categorieën 'fraude' en 'geen fraude', zijn jaarverslagen nodig waarvan bekend is dat zij in een van deze categorieën vallen. Welke jaarverslagen fraude bevatten is bepaald op basis van nieuwsberichten en de Accounting and Auditing Enforcement Releases (AAER's) die gepubliceerd worden door de Securities and Exchange Commission (SEC), de Amerikaanse beurswaakhond. Fraudezaken worden alleen geselecteerd uit nieuwsberichten wanneer deze worden beschreven in meerdere media en als onderzoek heeft aangetoond dat er sprake was van fraude. Zaken die nog worden onderzocht of waarbij de conclusie fraude niet kan worden getrokken worden niet meegenomen. De AAER's bevatten alle sancties van de SEC, niet alleen die gerelateerd aan fraude. Voor het bepalen van frauduleuze jaarverslagen zijn alleen de AAER's geselecteerd waarin het woord 'fraud' en een term die wijst op een jaarverslag, zoals '10-K', '20-F'[3] of 'annual report', voorkomen. Op basis van de informatie verstrekt in de nieuwsberichten en de AAER's zijn de jaarverslagen verzameld die in de dataset de categorie 'fraude' vormen.

Voor elk jaarverslag in de categorie 'fraude' zijn ten minste drie jaarverslagen verzameld van gelijksoortige bedrijven, waarvoor geen frauduleuze activiteiten bekend zijn. Wanneer fraude niet gedetecteerd is betekent dat niet dat er geen fraude heeft plaatsgevonden. In dit onderzoek wordt het principe 'onschuldig tot het tegendeel bewezen is' gevolgd. De 'fraude'- en 'geen fraude'-jaarverslagen worden gematcht op basis van het jaar waarop het jaarverslag betrekking heeft, de sector waarin het bedrijf werkzaam is en het aantal medewerkers als indicatie voor de omvang van het bedrijf. Vanwege de aanname dat er meer bedrijven niet betrokken zijn bij fraude dan wel, zijn voor elk frauduleus jaarverslag meerdere niet-frauduleuze jaarverslagen geselecteerd. Het totaal aantal jaarverslagen in de dataset is 1.727. Hiervan vallen er 402 in de categorie 'fraude' en 1.325 in de categorie 'geen fraude'. De geselecteerde jaarverslagen gaan over de periode van 1999 tot en met 2011. Door de goede vastlegging en openbaar maken van sancties en jaarverslagen door de SEC bestaat het merendeel van de dataset uit jaarverslagen van Amerikaanse en niet-Amerikaanse bedrijven die op Amerikaanse beurzen genoteerd zijn.

Het onderzoek beschreven in dit artikel wordt uitgevoerd met de management discussion and analysis-sectie (MD&A) uit de jaarverslagen. Dit is het meest gelezen deel van het jaarverslag (Li 2010). In jaarverslagen die gerapporteerd worden aan de SEC, de Form 10-K en Form 20-F, is de MD&A-sectie een apart gedefinieerde sectie. De overige jaarverslagen zijn 'free format', maar bevatten wel secties die de financiële resultaten van het afgelopen jaar beschrijven. Het handmatig extraheren van de MD&A-secties is tijdrovend. Wij hebben daarom een methode ontwikkeld die deze secties automatisch uit de 10-K en 20-F jaarverslagen haalt. Het algoritme herkent

het begin van de MD&A-sectie op basis van de eerste vier woorden van het kopje van de sectie. Ook de sectie volgend op de MD&A-sectie wordt op deze manier herkend zodat het algoritme weet waar de MD&A-sectie eindigt. Het algoritme houdt rekening met variaties die bestaan in de kopjes. Met deze methode is de MD&A-sectie van 96% van de jaarverslagen die zijn gerapporteerd op form 10-K of form 20-F geselecteerd. De MD&A-secties van de andere 10-K, 20-F en overige jaarverslagen zijn handmatig geëxtraheerd.

### 3.2 Feature extraction and selection

Een *machine learning*-model heeft een gestructureerde representatie van de data nodig om patronen in deze data te kunnen leren. Het omzetten van tekst naar een gestructureerde representatie die een computer kan begrijpen wordt *feature extraction and selection* genoemd. Er bestaat een verscheidenheid aan *features* die uit tekst kunnen worden bepaald. Voor het onderzoek beschreven in dit artikel zijn de *features* bepaald die in eerdere onderzoeken naar fraudedetectie in tekst en leugendetectie succesvol waren.

In essentie is tekst een aaneenschakeling van woorden. Het is daarom begrijpelijk dat de meest gebruikte *feature* in *text mining* gebaseerd is op het tellen van individuele woorden, '*word unigrams*' genoemd (Jurafsky and Martin 2000). Probleem hierbij is dat het aantal woorden in documenten vaak te groot is. Om dit te ondervangen wordt gebruik gemaakt van *feature selection*-stappen. Deze reduceren het aantal *features* zonder verlies van informatie. In dit onderzoek hebben we een statistische methode, de chi-kwadraat, gebruikt voor het selecteren van de mogelijk meest informatieve *features*. Initieel selecteren we de top 1.000 meest informatieve *features* als input voor de *machine learning*-algoritme beschreven in paragraaf 3.3. Deze top 1.000 wordt incrementeel verhoogd met de volgende top 1.000 meest informatieve *features* om te bepalen of het toevoegen van meer *features* het resultaat van het model verbetert. Op eenzelfde manier als de *word unigrams* kunnen groepjes van twee opeenvolgende woorden dienen als *features*. Dit worden '*word bigrams*' genoemd. Een combinatie van *word unigrams* en *word bigrams* is ook een mogelijkheid.

Naast het tellen van de individuele woorden bestaan er *features* die gebaseerd zijn op formules of het tellen van woordcategorieën. De formules beschrijven de teksten op verschillende manieren. Zo bestaat er een formule die de lexicale diversiteit berekent. De lexicale diversiteit drukt uit hoeveel verschillende woorden gebruikt worden in de tekst. De tekstcomplexiteit kan gemeten worden aan de hand van de formules die bestaan uit de gemiddelde zinslengte, het percentage lange zinnen en het percentage complexe woorden dat voorkomt in de tekst. Formules voor leesbaarheid van de tekst berekenen veelal hoeveel jaar onderwijs iemand nodig heeft om de tekst te kunnen begrijpen. De grammaticale *features* vatten de soorten woordgroepen die voorkomen in de tekst of de constructie van de zinnen samen. De woordgroepen bestaan uit woorden met grammaticale kenmerken die overeenkomen. Voorbeelden hiervan zijn werkwoorden en zelfstandig naamwoorden. Ten slotte is er een categorie met psychologische *features*. Sinds het begin van psychologisch onderzoek bestaat het idee dat mensen gedachten en intenties uitdrukken via taal (Tausczik and Pennebaker 2009). Pennebaker et al. (2007) hebben een programma, de '*Linguistic Inquiry and Word Count*' (LIWC) ontwikkeld die dergelijke gedachten en intenties mogelijk extraheert. De psychologische kenmerken die in onderzoeken naar leugendetectie succesvol waren zijn onder andere positief en negatief woordgebruik en woorden die duiden op angst of een mate van zekerheid. Op deze manier onderscheiden wij in totaal zes categorieën *features*, te weten 'woorden tellen', 'beschrijvend', 'tekstcomplexiteit', 'leesbaarheid', 'grammatica' en 'psychologisch'.

### 3.3 Machine learning

*Machine learning* is een *computer science*-methode en een breed onderzoeksveld binnen kunstmatige intelligentie dat zich bezig houdt met de ontwikkeling van algoritmen en technieken waarmee computers kunnen leren (Wikipedia). In dit onderzoek leert een *machine learning*-algoritme op basis van de *features* wanneer een jaarverslag aangemerkt kan worden als frauduleus of niet-frauduleus.

De ontwikkeling van een *machine learning*-model vereist twee datasets: een dataset voor het ontwikkelen van het model, de ontwikkelingsset, en een dataset waarop de prestaties van het model kunnen worden getest, de validatieset. Daarom wordt de totale dataset beschreven in paragraaf 3.1 willekeurig verdeeld in een ontwikkelingsset en een validatieset. De ontwikkelingsset omvat 70% van de gegevens, terwijl de resterende 30% bewaard wordt om de prestaties van het definitieve model te evalueren, nadat de ontwikkeling is voltooid.

Een *machine learning*-algoritme heeft *features* nodig als input. Voor dit onderzoek zijn op de ontwikkelingsset experimenten uitgevoerd met verschillende combinaties van de *features* zoals beschreven in paragraaf 3.2. Er ontstaan daardoor meerdere modellen. Het basismodel gebruikt alleen de *word unigrams* als input voor de *machine learning*-algoritmen. De andere categorieën *features,* 'beschrijvend', 'tekstcomplexiteit', 'leesbaarheid', 'grammatica' en 'psychologisch', worden hier om de beurt aan toegevoegd om vast te stellen of deze categorieën informatie toevoegen aan het basismodel dat alleen gebruik maakt van de *unigrams.* De beste resultaten in deze ontwikkelingsfase worden vervolgens getest op de validatieset.

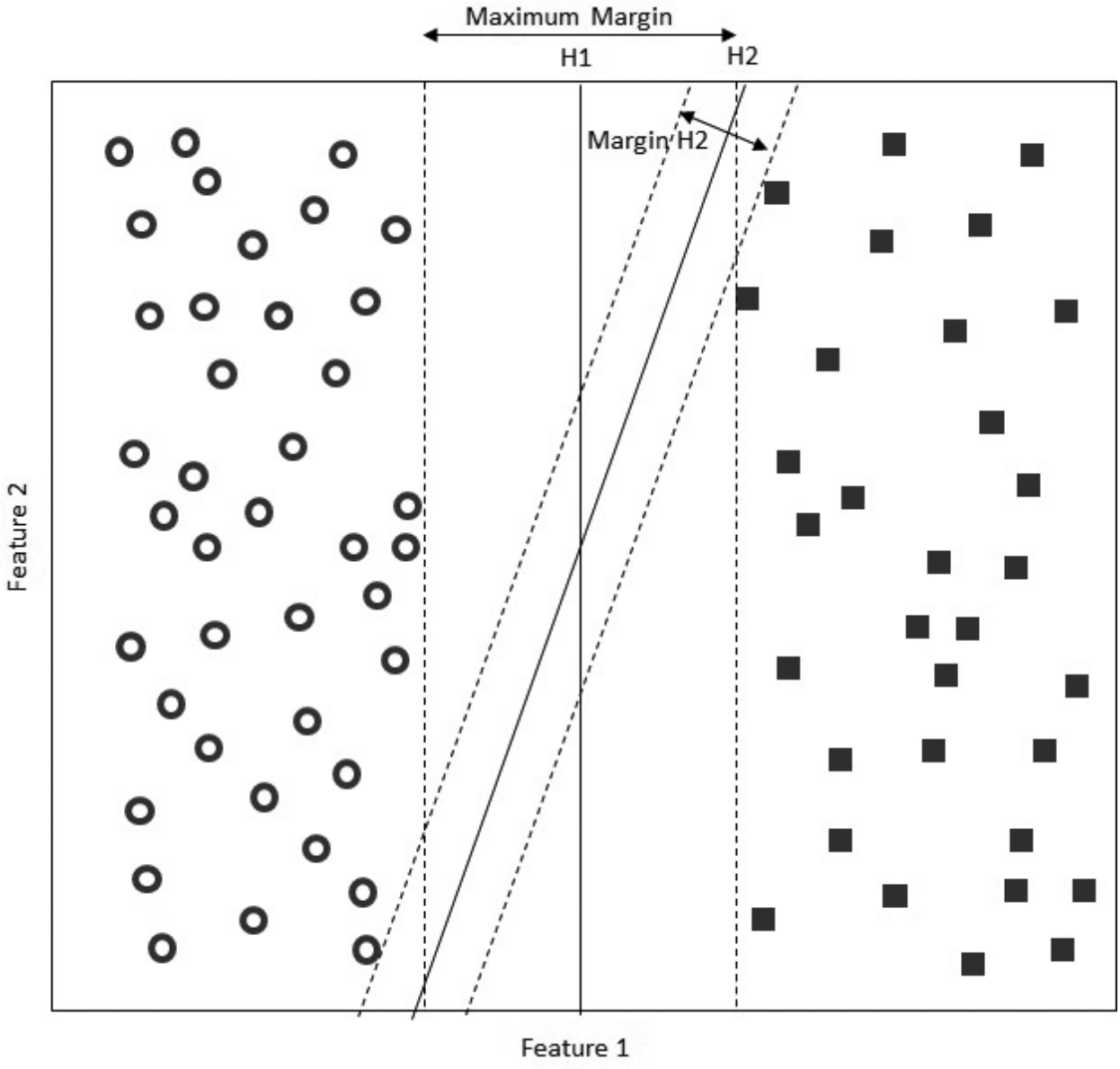
Er bestaat een verscheidenheid aan *machine learning*-algoritmen. Voor dit onderzoek gebruiken we twee *machine learning*-algoritmen die in eerdere text mining-onderzoeken succesvol waren, te weten *Naive Bayes* (NB) en *Support Vector Machine* (SVM) (Cecchini et al. 2010; Conway et al. 2009; Glancy and Yadav 2011;

Goel et al. 2010; He and Veldkamp 2012; Joachims 1998; Manning and Schütze 1999; Metsis et al. 2006; Purda and Skillicorn 2015). Beide algoritmen kiezen een andere benadering. Het NB-algoritme berekent de kansen dat een jaarverslag behoort tot de twee categorieën `fraude' en `geen fraude'. In de berekening van deze kansen gebruikt het algoritme de *features.* Het jaarverslag wordt vervolgens toegewezen aan de categorie met de hoogste kans. Het SVM-algoritme wijst elk jaarverslag als een punt toe in een wiskundige ruimte op basis van de *features.* De categorie waaraan een jaarverslag wordt toegewezen hangt af van de locatie van het jaarverslag in die ruimte. Een jaarverslag bevindt zich aan de kant van 'fraude' of aan de kant van 'geen fraude' van de door de SVM bepaalde scheidslijn. Voor meer informatie over *machine learning* en *machine learning*-algoritmen verwijzen we naar Tan et al. (2005); Joachims (1998); Manning and Schütze (1999).

# 4. Resultaten

Deze paragraaf beschrijft de resultaten van de *machine learning*-modellen. Ten eerste worden de prestatiematen toegelicht waarin de resultaten van de modellen worden uitgedrukt. Vervolgens worden de resultaten die zijn verkregen op de ontwikkelingsset gegeven. De *machine learning*-modellen die het beste presteerden op deze ontwikkelingsset, zijn vervolgens toegepast op de validatieset. De paragraaf eindigt met de resultaten die zijn behaald op deze validatieset. De betekenis van deze resultaten wordt verder besproken in paragraaf 5, de conclusie.

De prestaties van de modellen worden gemeten door middel van zes maten die uitdrukken hoe goed een *machine learning*-model de frauduleuze en niet-frauduleuze jaarverslagen kan detecteren. De meest ge-




**Afbeelding 1.** Voorbeeld van een eenvoudige Support Vector Machine (SVM) die twee categorieën van elkaar scheidt.

bruikte maat is nauwkeurigheid (*'accuracy'*), die het percentage jaarverslagen dat wordt toegekend aan de juiste categorie aangeeft. De tweede en derde maten zijn '*recall*' en '*precision*'. *Recall* geeft het percentage jaarverslagen dat toegewezen is aan de categorie 'geen fraude' waarbij ook daadwerkelijk sprake is van 'geen fraude'. *Precision* is het percentage jaarverslagen toegewezen aan de categorie 'fraude' dat ook daadwerkelijk frauduleus is. *Recall* en *precision* zijn maten die aangeven in hoeverre op de uitkomst van het model kan worden vertrouwd. Een vierde maat, de *F1-score*, combineert *recall* en *precision* als één waarde die de betrouwbaarheid van het model weergeeft. De vijfde maat, '*sensitivity*', berekent het percentage frauduleuze jaarverslagen dat door het model correct wordt aangemerkt als frauduleus. Op een vergelijkbare manier meet de zesde maat, '*specificity*', het percentage jaarverslagen in de categorie 'geen fraude' dat het model correct classificeert.

Tijdens het onderzoek zijn meerdere *machine learning*-modellen ontwikkeld die gebruik maken van verschillende combinaties van *features.* De modellen die gebruik maken van het tellen van woorden (*unigrams)* vormen de basis. Deze modellen zijn vervolgens uitgebreid met de andere features onderverdeeld in de categorieën 'beschrijvend', 'tekstcomplexiteit', 'leesbaarheid', 'grammatica' en 'psychologisch'. zoals genoemd in paragraaf 3.2. De beste resultaten op de ontwikkelingsset worden behaald door de modellen gebaseerd op *unigrams* en *bigrams* als *features*. Deze modellen maken dus alleen gebruik van het tellen van woorden. De andere *features* genoemd in paragraaf 3.2, zoals de *features* die de tekst beschrijven, de complexiteit en leesbaarheid meten of de grammaticale en psychologische informatie weergeven voegden geen informatie toe aan de modellen die alleen gebaseerd zijn op het tellen van woorden. De drie modellen die de beste resultaten behaalden tijdens de ontwikkeling zijn het NB-model met de top 10.000 *unigrams*, het SVM-model met de top 10.000 *unigrams* gecombineerd met de top 10.000 *bigrams* en een SVM-model met een top 30.000 van *unigrams* en *bigrams*. Deze drie modellen behaalden allen een *accuracy* van 89% op de ontwikkelingsset. De overige vijf prestatiematen varieerden voor de deze modellen. Het SVM-model met 30.000 *features* behaalde de hoogste '*specificity*', '*precision*' en *F1-score*, maar de laagste '*sensitivity*' en '*recall*'. Het NB-model behaalde de hoogste '*sensitivity*' en '*recall*, maar de laagste '*precision*' en *F1-score*.

De drie modellen die de beste resultaten behaalden op de ontwikkelingsset zijn toegepast op de validatieset. Tabel 1 toont de resultaten van deze modellen op de validatieset. Alle drie de modellen behalen een hoge *accuracy.* Zij zijn dus allen in staat om bijna 90% van de jaarverslagen toe te wijzen aan de juiste categorie, 'fraude' of 'geen fraude'. Het SVM-model met de top 10.000 *unigrams* en top 10.000 *bigrams* (in totaal 20.000 *features* en derhalve in de tabel aangeduid als 'SVM 20.000') behaalde de hoogste '*accuracy'* en F1-score. Het NB-model behaalde de hoogste '*sensitivity*' en '*recall*', maar de laagste '*precision*' en *F1-score*. Het SVM model met de 30.000 *features* scoorde het hoogst op '*specificity*' en '*precision*', maar het laagst op 'accuracy', *'sensitivity*' en '*recall*'. De volgende paragraaf gaat in op de interpretatie van deze resultaten voor de praktijk.

# 5. Conclusie

De resultaten van het onderzoek laten zien dat het mogelijk is om indicaties van fraude in jaarverslagen te herkennen door middel van tekstanalyse. Deze paragraaf gaat eerst dieper in op de interpretatie van de uitkomsten van de prestatiematen. Vervolgens wordt het ontbreken van de toegevoegde waarde van de *features* die gebaseerd zijn op het tellen van woordcategorieën en formules besproken.

Het SVM-model met 20.000 *unigrams* en *bigrams* laat met de hoogste scores op *accuracy*, *precision* en de *F1-score* het beste resultaat zien op de validatieset. Het NB-model met 10.000 *unigrams* scoort echter hoger op *sensitivity* en *recall*. Een hogere *sensitivity* betekent dat het model goed is in het detecteren van de frauduleuze jaarverslagen. Een hoge *recall* geeft aan dat, als het model een jaarverslag toewijst aan de categorie 'geen fraude', dit resultaat betrouwbaar is. Als we ervoor kiezen om niet verder te gaan met het onderzoeken van jaarverslagen die door het NB-model worden toegewezen aan de categorie 'geen fraude', dan zouden we weinig fraudegevallen missen. De *precision* van het NB-model is echter iets lager. Het model wijst jaarverslagen toe aan de fraudecategorie die niet frauduleus zijn. Wanneer de keuze wordt gemaakt om de jaarverslagen die toegewezen zijn aan de categorie 'fraude' verder te onderzoeken worden ook deze niet-frauduleuze jaarverslagen onderworpen aan verder onderzoek, wat extra kosten met zich meebrengt. Dit maakt het NB-model wellicht niet het meest kosteneffectief, maar omdat het model het minste aantal fraudegevallen mist kan deze worden beschouwd als het meest veilige model om op te vertrouwen. Het SVM-model met 30.000 *unigrams* en *bigrams* als input heeft de laagste *sensitivity* en detecteert dus het minste aantal fraudegevallen. De hoge mate van precisie laat echter zien dat als dit model een jaarverslag classificeert als frauduleus het model het hoogstwaarschijnlijk bij het juiste eind heeft. Geen van de *features* die gebaseerd zijn op het tellen van woordcategorieën of formules voegt informatie toe aan de NB- of SVM-modellen voor het detecteren van fraude in jaarverslagen van bedrijven wereldwijd, hoewel deze taalkundige kenmerken in onderzoeken naar fraude of leugendetectie relevant zijn bevonden. Er zijn meerdere redenen waarom deze kenmerken het resultaat van de *machine learning*-modellen die gebruik maken van de *unigrams* niet verbeteren. Ten eerste kijken de eer-

**Tabel 1.** Overzicht van de resultaten op de validatieset voor de drie modellen met de beste resultaten op de ontwikkelingsset.

| Model | Accuracy | Sensitivity | Specificity | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| NB 10.000 | 0,89 | 0,72 | 0,95 | 0,81 | 0,92 | 0,86 |
| SVM 20.000 | 0,90 | 0,60 | 0,99 | 0,95 | 0,89 | 0,92 |
| SVM 30.000 | 0,87 | 0,45 | 1,00 | 0,98 | 0,86 | 0,91 |

dere onderzoeken naar statistische verschillen. Echter, *features* met statistische significantie hoeven niet relevant te zijn voor een *machine learning*-algoritme (Zhou et al. 2004). Ten tweede bestaat er ambiguïteit in de relatie tussen de aanwezigheid van fraude of leugens en de taalkundige *features*. Voor een deel van de *features* vinden onderzoekers dat een verhoogde aanwezigheid van bepaalde *features* op bedrog wijst, terwijl andere onderzoekers het tegenovergestelde concluderen. Daarnaast is, in tegenstelling tot veel van de onderzoeken naar leugendetectie, het huidige onderzoek uitgevoerd met formele documenten. Het totaal aantal gebruikte taalkundige *features* is 72, wat een beperkt aantal is ten opzichte van de 10.000 *unigrams*. De taalkundige kenmerken zijn samenvattend, zij tellen woordgroepen. De *unigrams* kijken naar individuele woorden en zijn daardoor mogelijk in staat om subtielere verschillen in taalgebruik te vinden.

# 6. Ten slotte

Wij willen er op wijzen dat voor geen van de modellen bekend is wat de door het *machine learning*-algoritme gevonden patronen en woorden zijn om te bepalen of een jaarverslag wordt geclassificeerd als ‘fraude’ of als ‘geen fraude’. Benadrukt moet worden dat de modellen beschreven in dit artikel geen definitieve conclusie geven in de vaststelling van fraude. De classificatiebeslissing van een model moet als, ‘red flag’, worden gebruikt om, in geval van fraude, een bedrijf en zijn jaarverslag verder te onderzoeken. De beschreven methode leent zich uitstekend voor het screenen van jaarverslagen in kort tijdsbestek. De ontwikkelde methode is geschikt om tools te ontwikkelen die in de praktijk snel inzetbaar zijn, ook indien jaarverslagen niet direct digitaal beschikbaar zijn. Kosten, verbonden aan het beoordelen van jaarverslagen, worden daardoor gereduceerd.


- **Dr. Marcia Fissette** promoveerde aan de Universiteit Twente terwijl zij werkzaam was bij KPMG Forensic Technology. Het promotieonderzoek is gericht op fraudedetectie in jaarverslagen door middel van tekstanalyse.
- **Prof. dr. Bernard Veldkamp** is hoogleraar onderzoeksmethodologie en data analytics aan de Universiteit Twente. Hij specialiseert zich in social data analytics en computerized assessment.
- **Prof. dr. T. de Vries** is verbonden aan de Universiteit van Twente. Zijn belangstelling richt zich op het gebruik van nieuwe datatechnieken om fraude en gezondheidsaandoeningen te detecteren.


# Noten

1. Dit artikel is gebaseerd op het proefschrift ‘Text mining to detect indications of fraud in annual reports worldwide’ geschreven door Marcia Fissette onder supervisie van Bernard Veldkamp en Theo de Vries. Het volledig proefschrift is te vinden op https://research.utwente.nl/en/publications/text-mining-to-detect-indications-of-fraud-in-annual-reports-worl.
2. Form10-K is het formulier waarop beursgenoteerde Amerikaanse bedrijven het jaarverslag rapporteren aan de beurswaakhond, de Securities and Exchange Commisson (SEC).
3. Form 20-F is het formulier waarop niet-Amerikaanse bedrijven die genoteerd zijn aan een Amerikaanse beurs het jaarverslag rapporteren aan de beurswaakhond, de Securities and Exchange Commission (SEC).

# Literatuur


- Agostini M, Favero G (2013) Accounting fraud, business failure and creative auditing: A micro-analysis of the strange case of Sunbeam Corp. Working Papers 12, Department of Management, Università Ca’ Foscari Venezia. SSRN e-Library. https://ssrn.com/abstract=2149552
- Balakrishnan R, Qiu XY, Srinivasan P (2010) On the predictive ability of narrative disclosures in annual reports. European Journal of Operational Research 202(3): 789–801. https://doi.org/10.1016/j.ejor.2009.06.023
- Beasley MS, Carcello JV, Hermanson DR, Neal TL (2010) Fraudulent financial reporting: 1998–2007: An analysis of US public companies. The Committee of Sponsoring Organizations of the Treadway Commission (COSO). https://www.coso.org/Documents/COSO-Fraud-Study-2010-001.pdf
- Beattie V, Dhanani A, Jones MJ (2008) Investigating presentational change in U.K. annual reports: A longitudinal perspective. Journal of Business Communication 45(2): 181–222. https://doi.org/10.1177/0021943607313993

- Bell T, Carcello J (2000) A decision aid for assessing the likelihood of fraudulent financial reporting. Auditing: A Journal of Practice and Theory 19(1): 169–178. https://doi.org/10.2308/aud.2000.19.1.169
- Beneish MD (1997) Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance. Journal of Accounting and Public Policy 16(3): 271–309. https://doi.org/10.1016/S0278-4254(97)00023-9
- Burgoon J, Blair J, Qin T, Nunamaker Jr. JF (2003) Detecting deception through linguistic analysis. In: Chen H., Miranda R, Zeng D, Demchak C, Schroeder J, Madhusudan T (eds.), Intelligence and Security Informatics, volume 2665 of Lecture Notes in Computer Science. Springer (Berlin Heidelberg): 91–101. https://doi.org/10.1007/3-540-44853-5_7
- Cecchini M, Aytug H, Koehler GJ, Pathak P (2010) Making words work: Using financial text as a predictor of financial events. Decision Support Systems 50(1): 164–175. https://doi.org/10.1016/j.dss.2010.07.012
- Clatworthy MA, Jones MJ (2003) Financial reporting of good news and bad news: evidence from accounting narratives. Accounting and Business Research 33(3): 171–185. https://doi.org/10.1080/00014788.2003.9729645
- Conway M, Doan S, Kawazoe A, Collier N (2009) Classifying disease outbreak reports using n-grams and semantic features. International Journal of Medical Informatics 78(12): 47–58. https://doi.org/10.1016/j.ijmedinf.2009.03.010
- DePaulo BM, Lindsay JJ, Malone BE, Muhlenbruck L, Charlton, K, Cooper H (2003). Cues to deception. Psychological Bulletin 129(1): 74–118. https://doi.org/10.1037/0033-2909.129.1.74
- Fanning K, Cogger K (1998) Neural network detection of management fraud using published financial data. Intelligent Systems in Accounting, Finance and Management 7(1): 21–41. https://doi.org/10.1002/(SICI)1099-1174(199803)7:1<21::AID-ISAF138>3.0.CO;2-K
- Gee J, Button M, Brooks G (2017) The financial cost of fraud: what data from around the world shows. https://www.croweclarkwhitehill.co.uk/wp-content/uploads/sites/2/2017/02/crowe-the-financial-cost-of-fraud-2017.pdf
- Glancy FH, Yadav SB (2011) A computational model for financial reporting fraud detection. Decision Support Systems 50: 595–601. https://doi.org/10.1016/j.dss.2010.08.010
- Goel S, Gangolly J (2012) Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. Intelligent Systems in Accounting, Finance and Management 19(2): 75–89. https://doi.org/10.1002/isaf.1326
- Goel S, Gangolly J, Faerman SR, Uzuner O (2010) Can linguistic predictors detect fraudulent financial filings? Journal of Emerging Technologies in Accounting 7: 25–46. https://doi.org/10.2308/jeta.2010.7.1.25
- Grove H, Basilisco E (2008) Fraudulent financial reporting detection: Key ratios plus corporate governance factors. International Studies of Management & Organization 38(3): 10–42. https://doi.org/10.2753/IMO0020-8825380301
- Hájek P, Olej V (2013) Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In: Iliadis L, Papadopoulos H, Jayne C (eds.) Engineering Applications of Neural Networks. Springer (Berlin, Heidelberg): 1–10. https://doi.org/10.1007/978-3-642-41016-1_1
- He Q, Veldkamp DB (2012) Classifying unstructured textual data using the product score model: an alternative text mining algorithm. In: Eggen T, Veldkamp B (eds.), Psychometrics in practice at RCEC. RCEC (Enschede): 47–62. https://doi.org/10.3990/3.9789036533744.ch5
- Hoogs B, Kiehl T, LaComb C, Senturk D (2007) A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud. Intelligent Systems in Accounting, Finance and Management 15(1–2): 41–56. https://doi.org/10.1002/isaf.284
- Huang S, Tsaih R, Lin W (2012) Unsupervised neural networks approach for understanding fraudulent financial reporting. Industrial Management & Data Systems 112(2): 224–244. https://doi.org/10.1108/02635571211204272
- Humpherys SL, Moffitt KC, Burns MB, Burgoon JK, Felix WF (2011) Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems 50(3): 585–594. https://doi.org/10.1016/j.dss.2010.08.009
- Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. In: Nédellec C, Rouveirol C (eds.). Machine Learning ECML-98. Springer (Berlin, Heidelberg): 137–142. https://doi.org/10.1007/BFb0026683
- Jurafsky D, Martin JH (2000) Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall PTR (Upper Saddle River, NJ, USA), 2nd edition.
- Kaminski KA, Wetzel TS, Guan L (2004) Can financial ratios detect fraudulent financial reporting. Managerial Auditing Journal 19(1): 15–28. https://doi.org/10.1108/02686900410509802
- Kirkos E, Spathis C, Manolopoulos Y (2007) Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications 32(4): 995–1003. https://doi.org/10.1016/j.eswa.2006.02.016
- Lee T (1994) The changing form of the corporate annual report. The Accounting Historians Journal 21(1): 215–232. https://doi.org/10.2308/0148-4184.21.1.215
- Li F (2006) Do stock market investors understand the risk sentiment of corporate annual reports? SSRN e-Library. https://ssrn.com/abstract=898181
- Li F (2010) The information content of forward-looking statements in corporate filings - A naive Bayesian machine learning approach. Journal of Accounting Research 48(5): 1049–1102. https://doi.org/10.1111/j.1475-679X.2010.00382.x
- Li F, Lundholm RJ, Minnis M (2011) A new measure of competition bases on 10-K filings: Derivations and implications for financial statement analysis. Social Science Research Network Paper Series.
- Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT Press (Cambridge, MA, USA).
- Metsis V, Androutsopoulos I, Paliouras G (2006) Spam filtering with Naive Bayes - which Naive Bayes? In: CEAS 2006 - Third Conference on Email and Anti-Spam, July 27–28, 2006, Mountain View, California USA. http://www2.aueb.gr/users/ion/docs/ceas2006_paper.pdf
- Ndofor HA, Wesley C, Priem RL (2015) Providing CEOs with opportunities to cheat. The effects of complexity-based information asymmetries on financial reporting fraud. Journal of Management 41(6): 1774–1797. https://doi.org/10.1177/0149206312471395
- Newman ML, Pennebaker JW, Berry DS, Richards JM (2003) Lying words: Predicting deception from linguistic styles. Person-

ality and Social Psychology Bulletin 29(5): 665–675. https://doi.org/10.1177/0146167203029005010

- Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ (2007) The development and psychometric properties of LIWC2007. LIWC.net (Austin, Texas, USA). http://www.liwc.net/LIWC2007LanguageManual.pdf
- Pennebaker JW, Mehl MR, Niederhoffer KG (2003) Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology 54(1): 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041
- Perols, J (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing: A Journal of Practice & Theory 30(2): 19–50.
- Persons OS (1995) Using financial statement data to identify factors associated with fraudulent financial reporting. Journal of Applied Business Research 11(3): 38–46. https://clutejournals.com/index.php/JABR/article/view/5858/5936
- Purda L, Skillicorn D (2015) Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. Contemporary Accounting Research 32(3): 1193–1223. https://doi.org/10.1111/1911-3846.12089
- Purda LD, Skillicorn D (2010) Reading between the lines: Detecting fraud from the language of fnancial reports. SSRN e-Library. http://ssrn.com/abstract=1670832
- Ravisankar P, Ravi V, Raghava Rao G, Bose I (2011). Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems 50(2): 491–500. https://doi.org/10.1016/j.dss.2010.11.006
- Rezaee Z (2005) Causes, consequences, and deterence of financial statement fraud. Critical Perspectives on Accounting 16(3): 277–298.
- Skillicorn DB, Purda LD (2012) Detecting fraud in financial reports. In: 2012 European Intelligence and Security Informatics Conference, EISIC 2012, Odense, Denmark, August 22–24, 2012, pages 7–13. https://ieeexplore.ieee.org/document/6298880/
- Smith M, Taffler RJ (1999) The chairman's statement - a content analysis of discretionary narrative disclosures. Accounting, Auditing & Accountability Journal 13(5): 624–647. https://doi.org/10.1108/09513570010353738
- Spathis C, Doumpos M, Zopounidis C (2002) Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques. European Accounting Review 11(3): 509–535. https://doi.org/10.1080/0963818022000000966
- Tan P-N, Steinbach M, Kumar V (2005) Introduction to data mining. Addison-Wesley Longman Publishing Co., Inc. (Boston, MA, USA), first edition.
- Tausczik YR, Pennebaker JW (2009) The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology 29(1): 24–54. https://doi.org/10.1177/0261927X09351676
- Wang B, Wang X (2012) Deceptive financial reporting detection: hierarchical clustering approach based on linguistic features. Procedia Engineering 29: 3392–3396. https://doi.org/10.1016/j.proeng.2012.01.500
- Wang I, Radich R, Fargher N (2011) An analysis of financial statement fraud at the audit assertion level. Technical report, Working Paper. https://business.uow.edu.au/content/groups/public/@web/@commerce/@econ/documents/doc/uow120444.pdf
- Zhou L, Burgoon JK, Twitchell, DP, Qin T, Nunamaker Jr JF (2004). A comparison of classification methods for predicting deception in computer-mediated communication. Journal of Management Information Systems 20(4):139–166. https://doi.org/10.1080/07421222.2004.11045779