Univerzita Pardubice

Fakulta ekonomicko-správní

Analýza sentimentu v textu výročních zpráv

Bc. Jana Boháčová

Diplomová práce
2015

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: **2014/2015**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Jana Boháčová**
Osobní číslo:	**E13482**
Studijní program:	**N6208 Ekonomika a management**
Studijní obor:	**Ekonomika a management podniku**
Název tématu:	**Analýza sentimentu v textu výročních zpráv**
Zadávající katedra:	**Ústav podnikové ekonomiky a managementu**

Zásady pro vypracování:

Cílem práce je shrnout současné možnosti analýzy sentimentu, provést sběr dat, provést analýzu sentimentu ve výročních zprávách a analyzovat závislosti mezi sentimentem (pozitivním a negativním) a vývojem na kapitálových trzích.

Osnova:
- Analýza sentimentu.
- Sběr a zpracování dat.
- Analýza sentimentu ve výročních zprávách vybraných podniků.
- Posouzení závislosti mezi sentimentem a vývojem podniků na kapitálových trzích.

Rozsah grafických prací:

Rozsah pracovní zprávy: **cca 60 stran**

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

CECCHINI, M., AYTUG, H., et al. Making Words Work: Using Financial Text as a Predictor of Financial Events. Decision Support Systems. 2010, roč. 50, č. 1, s. 164-175. ISSN 1873-5797.
LIU, B. Sentiment Analysis and Opinion Mining. 1. vyd. San Rafael: Morgan & Claypool Publishers, 2012. 180 s. ISBN 978-1608458844.
LOUGHRAN, T., MCDONALD, B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance. 2011, roč. 66, č. 1, s. 35-65. ISSN 1540-6261.
MINER, G., ELDER, J., et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. 1. vyd. Waltham, MA: Academic Press, 2012. ISBN 978-0123869791.
WEISS, S. M., INDURKHYA, N., ZHANG, T. Fundamentals of Predictive Text Mining. 1. vyd. New York: Springer, 2010. 226 s. ISBN 978-1849962254.

Vedoucí diplomové práce: **doc. Ing. Petr Hájek, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání diplomové práce: **29. září 2014**

Termín odevzdání diplomové práce: **30. dubna 2015**

doc. Ing. Renáta Myšková, Ph.D.
děkanka

L.S.

doc. Ing. Marcela Kožená, Ph.D.
vedoucí ústavu

V Pardubicích dne 29. září 2014

PROHLÁŠENÍ

Prohlašuji, že jsem tuto práci vypracovala samostatně. Veškeré literární prameny a informace, které jsem v práci využila, jsou uvedeny v seznamu použité literatury.

Byla jsem seznámena s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako Školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně Univerzity Pardubice.

V Pardubicích dne 30. 4. 2015 Bc. Jana Boháčová

PODĚKOVÁNÍ:

Tímto bych ráda poděkovala svému vedoucímu práce doc. Ing. Petru Hájkovi, Ph.D. za jeho odbornou pomoc, cenné rady a poskytnuté materiály, které mi pomohly při zpracování diplomové práce.

Děkuji též své rodině a svým blízkým, kteří mi po dobu psaní této práce poskytovali podporu.

ANOTACE

Tato diplomová práce je zaměřena na analýzu sentimentu a její současné možnosti využití, především v oblasti financí. Po teoretickém vysvětlení analýzy sentimentu je proveden sběr a zpracování dat pro analýzu sentimentu ve výročních zprávách vybraných podniků. Následně je pomocí regresní analýzy posouzen vliv sentimentu ve zveřejněné výroční zprávě na vývoj ceny akcií podniků na kapitálových trzích.

KLÍČOVÁ SLOVA

analýza sentimentu, názory, slovník sentimentu, regresní analýza, faktorová analýza

TITLE

Sentiment Analysis in the Text of Annual Reports

ANNOTATION

This thesis is focused on the sentiment analysis and its current application opportunities especially in financial domain. After the theoretical explanation of the sentiment analysis, data are collected and processed for the analysis of sentiment in the annual reports of the selected companies. Subsequently the effect of the sentiment on the companies' stock price development on the capital markets is evaluated using regression analysis.

KEYWORDS

Sentiment Analysis, Opinions, Sentiment Lexicon, Regression Analysis, Factor Analysis

OBSAH

SEZNAM TABULEK

SEZNAM ILUSTRACÍ

SEZNAM ZKRATEK

CSČ	Celkový součet čtverců
EAT	Zisk po zdanění (Earnings After Taxes)
FI	Index příznivosti (Favorability Index)
IE	Extrakce informací (Information Extraction)
IR	Vyhledávání a získávání informací (Search and Information Retrieval)
MD&A	Diskuze a analýza managementu (Management Discussion and Analysis)
NASDAQ	Americká akciová burza založená Asociací národních obchodníků s cennými papíry (National Association of Securities Dealers Automated Quotations)
NLP	Zpracování přirozeného jazyka (Natural Language Processing)
NSČ	Nevysvětlený součet čtverců
NYSE	New Yorská akciová burza (New York Stock Exchange)
P/B	Poměr tržní ceny akcie a její účetní hodnoty (Price-to-Book Value)
P/BV	Poměr tržní ceny akcie a její účetní hodnoty (Price-to-Book Value)
P/E	Poměr tržní ceny akcie a zisku na akcii (Price-Earnings Ratio)
ROA	Rentabilita celkových aktiv (Return on Assets)
ROE	Rentabilita vlastního kapitálu (Return on Equity)
SEC	Komise pro cenné papíry a burzy (Securities and Exchange Commission)
SVM	Metoda strojového učení (Support Vector Machines)
U.S.	Spojené státy (United States)
USD	Americký dolar (United States Dollar)
VAR	Vektorové autoregresní modely (Vector Autoregression)
VSČ	Vysvětlený součet čtverců

ÚVOD

Analýza sentimentu je jednou z metod analýzy textu, která slouží ke zjištění lidského sentimentu a názorů a také ke zjištění hodnocení produktů, služeb, podniků nebo jiných problémů. Tato metoda se tedy nezaměřuje na fakta, nýbrž na subjektivní názory. V posledních letech neustále vzrůstá význam internetu a spolu s tím roste i objem dat zveřejňovaných online. Lidé tak mají přístup k digitalizovaným informacím, které obsahují různé názory či sentiment. V souvislosti s tím je po roce 2000 analýze sentimentu věnována čím dál větší pozornost. Z oblasti informatiky se analýza sentimentu rozšířila i do oblasti managementu a sociálních věd.

Tato diplomová práce je zaměřena na analýzu sentimentu a její využití především v oblasti financí. **Cílem této práce je shrnout současné možnosti analýzy sentimentu, provést sběr dat a analýzu sentimentu ve výročních zprávách podniků a analyzovat závislosti mezi sentimentem a vývojem podniků na kapitálových trzích.** Vývoj podniků na kapitálových trzích je hodnocen na základě procentní změny ceny akcií jednotlivých podniků. Práce je rozdělena na čtyři hlavní kapitoly. První část práce je věnována dosavadním teoretickým poznatkům, druhá část je praktická.

První kapitola se věnuje analýze sentimentu obecně – obsahuje její vysvětlení, účel a zařazení. Následuje popis a dělení názorů, které jsou pro analýzu sentimentu nezbytné. Stejně tak je pro analýzu sentimentu nezbytný i slovník sentimentu, jehož vysvětlení a proces jeho vytváření jsou zde také popsány. Kapitola dále obsahuje definici názorů, pomocí níž lze následně definovat cíl a klíčové úkoly analýzy sentimentu. Poslední část kapitoly je věnována vybraným problémům analýzy sentimentu – odhalování spamu a kvality recenzí.

Následuje kapitola zaměřená na analýzu sentimentu konkrétně v oblasti financí. Nejprve jsou zde popsány zdroje sentimentu pro oblast financí – sentiment vyjadřovaný podniky, sentiment vyjadřovaný v médiích a sentiment vyjadřovaný na internetu. Další část kapitoly je věnována ekonometrickým modelům, pomocí nichž lze zjistit, zda má sentiment dopad na vybrané ukazatele. Hlavní část druhé kapitoly je věnována vlivu sentimentu na různé proměnné či události. Pozornost je zde zaměřena především na vliv sentimentu na ceny a výnosy cenných papírů. Dále je zde popsána např. predikce finančních podvodů a bankrotů pomocí finanční analýzy.

Třetí kapitola popisuje sběr a přípravu dat. Jsou zde zjišťovány hodnoty sentimentu ve výročních zprávách vybraných podniků – amerických bank. Hodnoty sentimentu byly

zjišťovány dle dvou různých slovníků – prvním byl slovník pozitivních a negativních slov Elaine Henry, druhým slovník Loughrana a McDonalda s pozitivními, negativními, právnickými, neurčitými, slabě modálními a silně modálními slovy. Následně jsou zde popsány i ostatní vybrané determinanty vývoje ceny akcií, mezi které patří např. tržní kapitalizace, objem akcií obchodovaných na burze nebo ROE, P/E a P/B.

Poslední, čtvrtá, kapitola je zaměřena na predikci změny ceny akcie pomocí analýzy sentimentu. Pro zjištění vlivu sentimentu je v této práci vybrána jedna z nejpoužívanějších statistických metod – regresní analýza, která je v této kapitole i teoreticky popsána. Vzhledem k problémům s korelací proměnných je pro extrakci proměnných využita i faktorová analýza (ta je zde též stručně vysvětlena). Dále je zde popsán samotný výsledek regresní a faktorové analýzy – vliv sentimentu a ostatních faktorů na změnu ceny akcií. V závěru kapitoly jsou ověřeny předpoklady regresního modelu.

1 ANALÝZA SENTIMENTU

Slovník cizích slov [28] vysvětluje sentiment jako cit nebo citlivost, citovost. Analýza sentimentu, dle [18] nazývána též jako *opinion mining* (dolování názoru), se však zabývá rozborem nejenom lidského sentimentu a názorů, ale též přístupů a hodnocení, vztažených k produktům, službám, podnikům, ale i jednotlivcům, různým problémům i tématům.

Pojem analýza sentimentu se poprvé objevil pravděpodobně v roce 2003, i když práce na téma sentimentu a názorů se objevovaly i v dřívějších letech. Zájem o analýzu sentimentu se však objevil už v práci od Yoricka a Biena z roku 1983 [35]. Oblast analýzy sentimentu a dolování názorů se však stala aktivním místem především po roce 2000 a to hned z několika důvodů [18]. Jedním z nich je široké využití analýzy sentimentu v různých sférách. Díky této analýze se lze věnovat i náročným problémům, které ještě nikdy dříve rozebírány nebyly. Dalším důvodem je nepřeberné množství dat obsahujících názory na internetu a sociálních sítí. Vznik a rychlé rozšíření analýzy sentimentu souvisí právě se sociálními sítěmi.

1.1 Analýza textu

Za výrazy dolování textu a analýza textu se skrývají technologie pro analýzu a zpracování polo-strukturovaných a nestrukturovaných textů. Sjednocujícím prvkem těchto metod je potřeba „proměnit text na čísla“ [22]. Díky tomu je možné na velké množství dat aplikovat analytické algoritmy.

V současné době je možné využít k dolování textu několik způsobů. Dle [22] existuje sedm praktických oblastí analýzy textu, které se vzájemně překrývají. Do těchto oblastí patří následující:

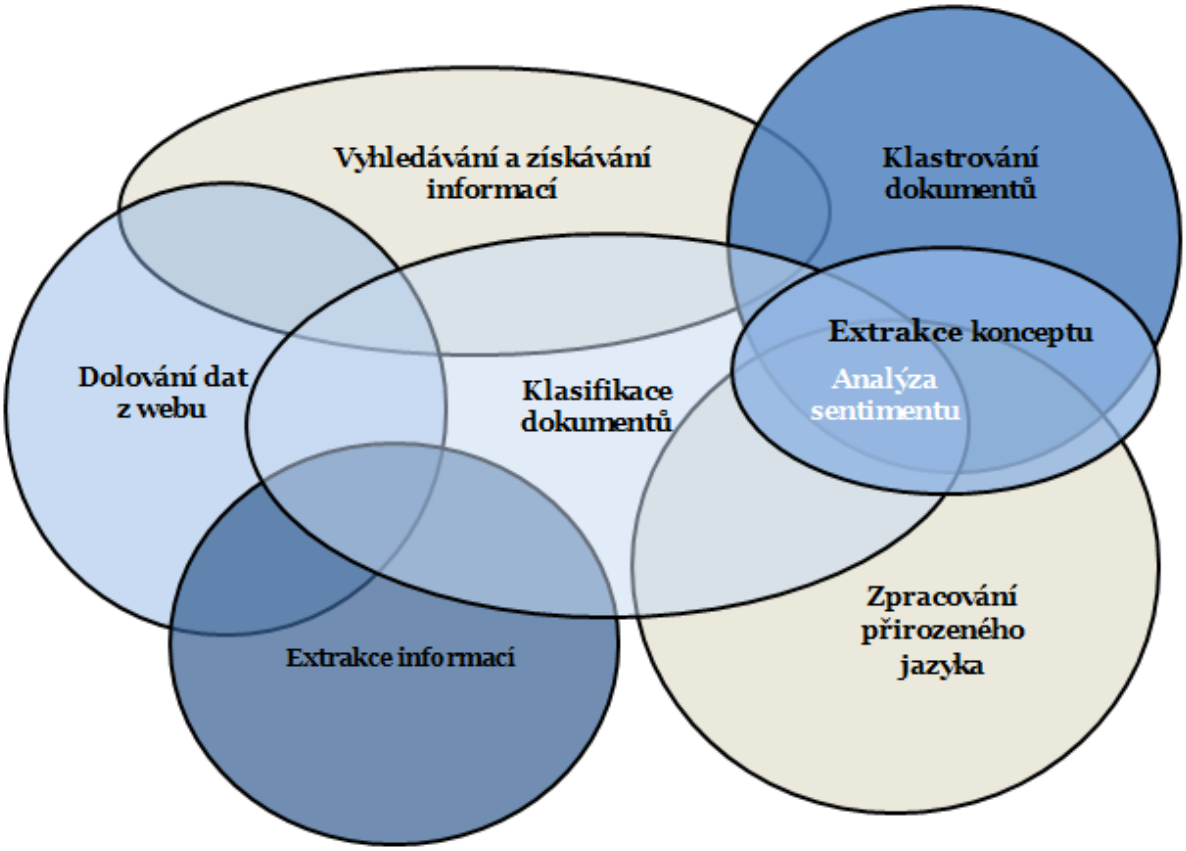
- *Search and Information Retrieval (IR)* – vyhledávání a získávání informací,
- *Document Clustering* – klastrování (shlukování) dokumentů,
- *Document Classification* – klasifikace dokumentů,
- *Web Mining* – dolování dat z webu,
- *Information Extraction (IE)* – extrakce informací,
- *Natural Language Processing (NLP*) – zpracování přirozeného jazyka,
- *Concept Extraction* – extrakce konceptu.

První z těchto sedmi metod je **vyhledávání a získávání informací** (IR). Tato metoda zahrnuje indexování, hledání a získávání dokumentů z rozsáhlých textových databází pomocí klíčových slov [22]. Se vzestupem internetových vyhledávačů (jako např. *Google*) se vyhledávání a získávání informací dostalo do podvědomí většiny lidí. Druhou metodou je **klastrování dokumentů**, které využívá algoritmů z dolování dat. Pomocí nich seskupuje podobné dokumenty do klastrů (shluků). **Klasifikace dokumentů** přiřazuje známé třídy nezatříděným dokumentům s využitím modelu, který se učil na zatříděných dokumentech.

Čtvrtou metodou je **dolování dat z webu**, které se týká dolování dat a textu na internetu, se specifickým zaměřením na propojenost webu [22]. Webové dokumenty jsou běžně strukturované a obsahují hypertextové odkazy, čímž se liší od běžných textů. Dolování dat z webu je stále se rozvíjející se oblast, která vychází z vyspělé technologie klasifikace dokumentů a porozumění přirozenému jazyku. **Extrakce informací** (IE) je jedna z více vyspělých metod v oblasti analýzy textu. Cílem této metody je vytvořit (extrahovat) strukturovaná data z nestrukturovaného textu. Začátečníci musí pro využití této metody vyvinout značné úsilí, protože vyžaduje specializované algoritmy a software.

Zpracování přirozeného jazyka (NLP) má poměrně dlouhou historii jak v lingvistice, tak v informatice. V poslední době se zaměření NLP posunulo blíže k oblasti dolování textu [22]. Zpracování přirozeného jazyka je výkonný nástroj pro poskytnutí vstupních proměnných pro dolování textu. Poslední metodou je **extrakce konceptu**, která se dá považovat za nejjednodušší a zároveň nejtěžší metodu. Tato metoda je zaměřena na seskupování slov a frází do sémanticky podobných skupin. Problémem této metody je fakt, že automatické systémy mají problém porozumět významu textu. Výchozí automatizované práce v kombinaci s lidským porozuměním však mohou vést k lepším výsledkům než samotná práce stroje nebo člověka.

Detailnější rozbor metod nabízí literatura, např. [22]. Jak lze spatřit na následujícím Obrázku 1, analýza sentimentu se nachází v průniku čtyř z těchto sedmi metod.



Obrázek 1: Pozice analýzy sentimentu vůči metodám zpracování textu

Zdroj: vlastní zpracování dle [22]

1.2 Názory

Názory lidí jsou klíčové pro oblast rozhodování. Podniky potřebují zjistit názory svých stávajících nebo potenciálních zákazníků na své produkty a služby, naopak zákazníky zajímá, zda už si produkt někdo zakoupil a jak jej hodnotí. Nejjednodušším způsobem zjištění názoru je prosté přímé zeptání se. Podniky obvykle provádějí průzkumy trhu, které si mohou zajistit samy nebo využít zprostředkovatele. Díky výše zmíněnému rozmachu sociálních sítí (především stránek typu *Twitter*, internetových blogů, různých diskuzních portálů apod.) se však v současné době mohou využívat i jiné možnosti jako například analýza sentimentu zákazníků [18].

1.2.1 Typy názorů

Názory mohou být rozděleny dvěma způsoby – na běžné a srovnávací nebo na explicitní a implicitní [18].

Běžný názor (většinou pouze názor), vyjadřuje názor pouze na jednu věc (produkt, službu apod.). Má dva hlavní podtypy – přímý a nepřímý názor. Přímý názor odkazuje přímo

na předmět, kdežto nepřímý je vyjádřen skrz jiné entity. Většina současných aktivit je zaměřena na přímé názory, se kterými se lépe pracuje. **Srovnávací názor** porovnává více věcí na základě nějakého jejich společného hlediska. Zkoumají se podobnosti nebo naopak odlišnosti dvou a více entit a také preference držitele názoru.

Explicitní názor (výslovný, zřetelný, zjevný) [28] je subjektivní sdělení, které může vyjadřovat běžný nebo srovnávací názor, např.: „Coca-cola chutná skvěle.“, „Coca-cola chutná lépe než Kofola.“ **Implicitní názor** (zahrnutý, obsažený) je objektivní sdělení, které obsahuje běžný nebo srovnávací názor [28]. Může vyjadřovat žádoucí, ale i nežádoucí fakta, jako např.: „Tato pračka spotřebuje hodně vody.“, „Baterie telefonu Nokia vydrží déle než Samsung.“.

1.2.2 Hledání a získávání názorů

Běžně vyhledávané názory na internetu jsou **veřejné hromadné názory** na určitou entitu (názor zákazníků na určitý produkt nebo názor lidí na politického kandidáta) nebo **názory jednotlivce nebo organizace** na určitou entitu nebo její aspekt (názor prezidenta na zákaz kouření v restauraci) [18]. V prvním případě stačí vyhledávat pouze entitu a/nebo její aspekt, v druhém je třeba vyhledávat entitu i spolu s držitelem názoru.

Stejně jako tradiční vyhledávání na internetu má i vyhledávání názoru dva hlavní úkoly a dva pod-úkoly [18]:

1. získat potřebné dokumenty nebo věty podle požadavků uživatele,
    a. najít dokumenty nebo věty, které odpovídají žádosti (tento pod-úkol je možné provést v tradičním internetovém prohledávání),
    b. určit, zda dokumenty nebo věty vyjadřují názor na žádané téma, a zda je tento názor pozitivní nebo negativní (v tomto případě je třeba použít analýzy sentimentu),
2. ohodnotit získané dokumenty nebo věty.

1.2.3 Shrnutí názorů

Na rozdíl od faktických informací jsou názory přirozeně subjektivní. Jeden názor od jednoho držitele názoru proto nemá smysl a je třeba analyzovat informace od velkého množství lidí. V tomto případě je vhodné využít nějakého souhrnu názorů – například strukturovaného souhrnu nebo krátkého textového souhrnu [18]. Klíčovými komponenty v takovém souhrnu jsou názory ohledně různých entit a jejich aspektů a kvantitativní hodnoty.

Obrázek 2 je příkladem strukturovaného souhrnu. Celkově hodnotí produkt (digitální fotoaparát) pozitivně 105 názorů, negativně pouze 12. Kvalita obrázků je označena jako pozitivní v 95 případech, negativní pouze v 10. Posledním hodnoceným aspektem je výdrž baterie, kde je 50 pozitivních hodnocení a 9 negativních.

Digitální fotoaparát **1:**

Aspekt: CELKOVÝ

Pozitivní:	105	<Jednotlivé větyrecenzí>
Negativní:	12	<Jednotlivé větyrecenzí>

Aspekt: Kvalita obrázků

Pozitivní:	95	<Jednotlivé větyrecenzí>
Negativní:	10	<Jednotlivé větyrecenzí>

Aspekt: Výdrž baterie

Pozitivní:	50	<Jednotlivé větyrecenzí>
Negativní:	9	<Jednotlivé větyrecenzí>

Obrázek 2: Stukturovaný souhrn názorů

Zdroj: vlastní zpracování dle [18]

1.2.4 Definování názoru

Aby mohl být určen cíl a klíčové úkoly analýzy sentimentu, je nejprve nutné definovat názor. Pro tento účel bude v této podkapitole využita fiktivní recenze dle [18] – autorem je John Smith, datum recenze 10. září 2011: „*Před šesti měsíci jsem si koupil fotoaparát Canon G12. Prostě ho zbožňuji. Kvalita obrázků je úžasná. Baterie vydrží také dlouho. Moje žena si však myslí, že je pro ni příliš těžký.*“

Výše zmíněná recenze vyjadřuje pozitivní, ale i negativní názor na produkt. Druhá věta obsahuje pozitivní názor na celý produkt, třetí a čtvrtá kladně hodnotí jeho konkrétní části. Poslední věta obsahuje negativní názor ohledně váhy fotoaparátu.

Hlavními komponenty názoru je terč (g) a sentiment (s), tedy

$$(g, s) \tag{1}$$

kde g je entita nebo aspekt entity, o kterém byl názor vyjádřen a s je pozitivní, negativní nebo neutrální sentiment nebo vyjadřuje skóre na bodové škále. Terčem hodnocení je ve druhé větě *fotoaparát Canon G12*, ve třetí větě pak *kvalita obrázků*.

Definice jako čtveřice

Recenze na počátku této kapitoly obsahuje názor dvou osob, které mohou být označeny jako zdroje názorů nebo držitelé názorů. Držitelem názoru je v druhé, třetí a čtvrté větě autor recenze (John Smith) a v poslední větě jeho žena. Důležitou částí recenze je též její datum, protože názory se s časem a trendy mění.

Názor tedy může být definován jako čtveřice, tedy

$$(g, s, h, t) \tag{2}$$

kde *g* (*target*) je terč názoru, *s* (*sentiment*) je sentiment, *h* (*sentiment holder*) je držitel názoru a *t* (*time*) je čas, kdy byl názor vyjádřen. Tuto definici však není jednoduché v praxi použít, protože plný popis terče se nemusí objevovat v jedné větě [18]. Ve výše zmíněné recenzi ve třetí větě je terč názoru „*kvalita obrázků fotoaparátu Canon G12*“, i když věta zmiňuje pouze „*kvalita obrázků*“.

Definice jako pětice

Výraz „*kvalita obrázků fotoaparátu Canon G12*“ může být rozdělen na entitu a atribut entity (*Canon G12* a *kvalita obrázků*). Entitou (*e*) může být produkt, služba, téma, problém, osoba, organizace nebo událost. Je definována jako *e: (T, W),* kde *T* je hierarchie částí, dílčích částí atd. a *W* je sada atributů entity. Každá část i dílčí část má také svou sadu atributů. Ve zmíněném příkladu je entitou *Canon G12*, který má sadu atributů – *kvalitu obrázků, velikost* a *váhu*. Fotoaparát má sadu částí, např. *čočky*, *hledáček* a *baterii*, která má vlastní atributy (*výdrž baterie* a *váhu baterie*).

Názor proto může být definován jako pětice [18], tedy

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l) \tag{3}$$

kde e_i je název entity, a_{ij} je aspekt e_i, s_{ijkl} je sentiment aspektu a_{ij} entity e_i, h_k je držitel názoru a t_l je čas, kdy byl názor vyjádřen držitelem názoru h_k. Sentiment s_{ijkl} může být opět pozitivní, negativní nebo neutrální nebo je vyjádřen skórem (např. od 1 do 5 hvězdiček), jak to často bývá u recenzí na internetu. Tato definice se týká pouze běžného názoru, pro srovnávací názor existuje odlišná definice.

Všech pět výše zmíněných komponentů musí být spolu v souladu. Názor s_{ijkl} musí být vytvořený držitelem h_k ohledně aspektu a_{ij} entity e_i v čase t_l. Žádný z nich by tak neměl chybět [18]. Problémem je například chybějící časový údaj, protože názor starý několik let

a názor současný se mohou velmi odlišovat. Taktéž chybějící držitel názoru může zkomplikovat analýzu.

1.3 Cíl a klíčové úkoly analýzy sentimentu

Pomocí definice zmíněné v předchozí podkapitole lze určit cíl a klíčové úkoly analýzy sentimentu, které jsou odvozeny od názorové pětice [18]. Prvním komponentem je **entita** a prvním úkolem je tak její získání. Po extrakci entit následuje jejich třídění. Každá entita má svou jedinečnou kategorii, v níž se nacházejí různé výrazy pro entitu. Příkladem kategorie entity může být *televize*, do této kategorie pak spadají výrazy *TV, telka, televize*.

Stejné třídění se týká i komponentu **aspekt entity**. Výrazy aspektu mohou být vyjádřeny přímo nebo nepřímo. V příkladu „*Kvalita obrázků je úžasná*," je aspekt *kvalita obrázků* vyjádřen přímo. Z věty „*Tento fotoaparát je drahý*," aspekt *cena* vyplývá, i když není přímo uveden.

Třetím komponentem pětice je **sentiment**, úkolem je zařadit sentiment aspektu jako pozitivní, negativní nebo neutrální. Stejně jako entity a aspekty také poslední komponenty (**držitel názoru** a **čas**) musí být získány a kategorizovány.

Analýza sentimentu dokumentu *d* se sestává z následujících šesti hlavních úkolů [18]:

1. **extrakce a kategorizace entity** – získání všech výrazů entit v dokumentu a kategorizace nebo vytvoření shluků pro synonymní výrazy entit, každý shluk vyjadřuje jedinečnou entitu e_i,
2. **extrakce a kategorizace aspektu** – získání všech výrazů aspektu a jejich kategorizace do shluků, každý shluk reprezentuje jedinečný aspekt a_{ij},
3. **extrakce a kategorizace držitele názoru** – stejně jako v předchozích dvou případech je třeba získat z textu držitele názoru a roztřídit je,
4. **extrakce a standardizace času** – získání časových údajů o jednotlivých názorech a standardizace časových formátů,
5. **klasifikace sentimentu aspektu** – rozhoduje se, zda názor na určitý aspekt a_{ij} je pozitivní, negativní nebo neutrální, nebo přiřazuje sentimentu číselné hodnocení,
6. **vytvoření názorové pětice** – vytvoření všech názorových komponentů (e_i, a_{ij}, s_{ijkl}, h_k, t_l) vyjádřených v dokumentu *d*.

Analýza sentimentu může probíhat na třech různých úrovních [18]:

- **úroveň dokumentu** - cílem na této úrovni je rozhodnout, zda celý dokument vyjadřuje pozitivní nebo negativní sentiment, pro tuto úroveň je vhodné použít dokument, který pojednává pouze o jedné věci (například jednom produktu nebo jednom podniku),
- **úroveň věty** - týká se pouze jedné konkrétní věty, rozhoduje se, zda je věta pozitivní, negativní či neutrální (bez názoru),
- **úroveň aspektu a entity** – na rozdíl od předchozích úrovní zjišťuje, co přesně se lidem líbí či nelíbí, názor lidí se sestává ze sentimentu (pozitivního či negativního) a terče názoru (názor bez terče má pouze omezené využití).

Každý dokument je určen jako pozitivní nebo negativní podle frekvence jednotlivých sentimentálních slov. Celkový sentiment je určen poměrem pozitivních a negativních slov. Možný průběh výpočtu indexu příznivosti dle [34] zachycují následující body a rovnice:

1. pro každý dokument je třeba vypočíst:

$$S = \frac{p}{(p+n)} \tag{4}$$

   kde S je sentiment, p je suma všech frekvencí pozitivních slov a n je suma všech frekvencí negativních slov,

2. dokument je pozitivní, pokud

$$S > t \tag{5}$$

   kde t je určená hranice (*treshold*) - nejčastěji 50 %, jinak je dokument negativní,

3. zjistit index příznivosti *FI* (*favorability index*), *FI* je založeno na celkových výsledcích všech *S*:

$$FI = \frac{počet\ pozitivních\ dokumentů}{počet\ negativních\ dokumentů} \tag{6}$$

1.4 Slovník sentimentu

Indikátory sentimentu jsou pozitivní či negativní výrazy a fráze. Slova obsahující sentiment mohou být rozdělena na dva typy – základní a komparativní typ. Druhý typ obsahuje komparativ a superlativ různých přídavných jmen a příslovcí [18]. Mezi nejběžnější pozitivní výrazy patří *dobrý*, *skvělý*, *úžasný*, naopak *špatný*, *hrozný* a *příšerný* budou slova

s negativním sentimentem. Seznam těchto slov a frází se nazývá **slovník sentimentu** (*sentiment lexicon*). Jeden slovník může být vytvořen s pozitivními slovy a frázemi a druhý s negativními [34].

Hlavními nositeli sentimentu jsou tedy přídavná jména. Význam těchto slov je možné posoudit ze tří základních pohledů [3]:

1. **hodnocení** (pozitivní/negativní),
2. **potence** (slabý/silný účinek),
    a. distance (vztah autora k tématu),
    b. specifičnost (formulace jasná/nejasná),
    c. určitost (jistý/váhavý autor),
3. **intenzita** (emotivnost výpovědi).

Slovník sentimentu je pro analýzu sentimentu nutný, ale ne dostačují. V souvislosti s ním se objevují různé problémy [18]. Prvním problémem je případ, kdy sentimentální slovo neobsahuje žádný sentiment. Například v otázce *„Můžeš mi doporučit nějaký dobrý fotoaparát od Canonu?“* není vyjádřeno pozitivní hodnocení. Dalším problémem je sarkasmus, který se často vyskytuje spíš v politických debatách, než v hodnocení produktů či služeb.

I věty bez sentimentálních slov však mohou vyjadřovat určitý sentiment. Věta *„Tahle pračka spotřebuje hodně vody,“* obsahuje jednoznačně negativní sdělení. Kromě zmíněného sarkasmu musí být slovník schopen postihnout též slang, idiomy nebo v současné době i grafické vyjádření emocí – to se týká např. „*smajlíků“*, označení *palec nahoru/dolů* a dalších [3].

Běžně využívaným zdrojem pro klasifikaci slov je *Harvardský psychosociologický slovník* (*Harvard Psychosociological Dictionary*), obzvláště *Harvard-IV-4 TagNeg (H4N)* [19]. Pozitivní stránkou tohoto seznamu je, že jeho obsah je mimo kontrolu badatele. Uživatel si tak nemůže vybrat, které slovo má negativní význam, je to přesně dáno. Anglická slova však mohou mít více významů, stejné slovo může v jedné oblasti znamenat něco jiného, než v oblasti druhé. Důležitou otázkou se pak stává, zda seznam slov sestavený pro psychologii a sociologii může stejně úspěšně posloužit pro oblast finanční.

Vytvoření slovníku

Slovník sentimentu může být vytvořen třemi způsoby – manuálním přístupem, přístupem založeném na slovníku a přístupem založeným na korpusu [18]. **Manuální přístup** je náročný na práci i čas a proto se využívá v kombinaci s automatickým přístupem jako jeho finální kontrola.

Prvním z automatických přístupů je **přístup založený na slovníku**. Nejprve se manuálně vytvoří sada sentimentálních slov (jader), u kterých je známa pozitivní nebo negativní orientace. Dále se algoritmem prohledá vybraný slovník pro synonyma a antonyma k původním vybraným slovům (jádrům) a nově nalezená slova se k nim přidají. Tento proces se dále opakuje, dokud nelze najít další nová slova. Po ukončení těchto automatických kroků následuje manuální kontrola.

Druhou automatickou metodou je **přístup založený na korpusu**, který má dvě hlavní možnosti. První možností je seznam slov se známým sentimentem (často s obecným účelem), ke kterým se získávají další sentimentální slova, a jejich orientace se zjišťuje podle korpusu. Druhou možností je s pomocí korpusu upravit slovník sentimentu s obecným účelem. Ačkoli přístup založený na korpusu může za pomocí velkého a různorodého korpusu vytvořit slovník sentimentu pro obecný účel, přístup založený na slovníku je obvykle efektivnější (protože slovník obsahuje všechna slova).

Při porovnání přístupu založeném na slovníku a metody strojového učení je přístup založený na slovníku pro uživatele jednodušší. V současné době pro uživatele existují i předem připravené programy [14] (např. *GI, DICTION*). Obecné slovníky se však nehodí pro textovou analýzu finančních souvislostí. Tento problém může být zmírněn využitím seznamu specifických slov (finančních), je však třeba nastavit správně váhu slov.

Dalším rozdílem metod je rychlost implementace. Zavádění strojového učení zabere více času a též nákladů, především kvůli manuálnímu ohodnocení „trénovací množiny“. Pro tento proces musí být vybráni vhodní lidé, např. rodilí mluvčí s ekonomickými a finančními zkušenostmi. Výhodou této metody je vyšší přesnost, někteří autoři uvádějí až o 30 % vyšší než byla metoda na základě slovníku [14].

1.5 Problémy analýzy sentimentu

Kromě zmíněných problémů v případě slovníku sentimentu se mohou objevit i jiné problémy komplikující analýzu sentimentu.

1.5.1 Odhalování spamu

Spam se nejčastěji objevuje v podobě nevyžádaných emailů a nežádoucích internetových stránek. V tomto případě se může jednat o nevyžádané informace jako například falešné názory a recenze [18]. S pozitivními recenzemi roste počet prodaných výrobků i dobrá image podniku a proto se mohou objevovat falešné negativní recenze např. od konkurence. Autoři těchto nevyžádaných spamů se nazývají *opinion spammers* (názoroví spameři) a jejich aktivitou je *opinion spamming* (spamování názorů).

Velmi nebezpečné může být spamování názorů ohledně sociálních nebo politických názorů, které může zmobilizovat davy. Spamování se v současné době se pořád zdokonaluje a propracovává a jeho odhalení je tak čím dál větší výzvou.

Problémem názorových spamů je obtížnost jejich odhalení. Na rozdíl od jiných druhů spamu (např. v emailech) je velmi těžké nebo i dokonce nemožné odlišit falešné recenze od pravých pouze jejich přečtením.

Dle [18] lze rozlišit tři typy spamových recenzí:

- **typ 1: falešné recenze** – tyto neupřímné recenze nejsou založeny na zkušenostech autora, jejich cílem je vychválit určitou entitu za účelem zvýšení prodeje nebo popularity nebo naopak zničit reputaci jiné entity (např. konkurence),
- **typ 2: recenze pouze o značkách** – tyto recenze mohou být upřímné, ale jsou považovány za spam, protože se netýkají konkrétních produktů nebo služeb, ale komentují pouze značky či výrobce těchto produktů,
- **typ 3: ne-recenze** – články, které nejsou recenzemi – mohou to být reklamy nebo další články bez názoru.

Druhý a třetí typ spamu se neobjevuje často a je relativně snadné jej odhalit. První typ (falešné recenze) může způsobovat různě vážné problémy, vždy záleží i na typu produktu. Velmi kvalitnímu produktu nemusí uškodit ani falešná negativní recenze, stejně tak jako falešná pozitivní recenze nemusí přilepšit nekvalitnímu produktu.

Falešné recenze mohou být psány různými typy lidí – rodinou a přáteli, zaměstnanci společnosti, konkurencí nebo i zákazníky, kterým byla přislíbena nějaká výhoda či sleva, pokud napíšou pozitivní recenzi.

Spamování může být prováděno **individuálně** nebo **skupinově** [18]. Spammer jednotlivec s nikým nespolupracuje, píše recenze samostatně a má jedinečný uživatelský účet. Skupina

spammerů se může objevovat ve dvou případech. V prvním z nich existuje skupina osob, která společně propaguje určitou entitu a/nebo se snaží zničit reputaci jiné entity. Jednotlivci v této skupině se mohou, ale i nemusí navzájem znát. Druhým případem je jednotlivec, který zaregistruje několik uživatelských účtů a chová se jako několik různých osob. Skupinové spamování může být velmi nebezpečné vzhledem k počtu členů a počtu recenzí. Může významně ovlivnit názor na produkt a zmást potencionální zákazníky především při uvedení výrobku na trh.

1.5.2 Kvalita recenzí

Kvalita recenzí souvisí s odhalováním spamů, ale obě témata jsou zároveň odlišná. Nekvalitní recenze sice není žádána, ale není to spam [18]. U každé recenze je třeba určit kvalitu, prospěšnost a užitečnost. V současné době na mnohých internetových stránkách existuje možnost ohodnotit jednotlivé recenze, zda pro zákazníka byly užitečné nebo ne.

Určení kvality recenzí je obvykle formulováno jako regresní problém. Vybraný model přiřazuje skóre kvality ke každé recenzi. Literatura uvádí různé typy modelů. Například Kim a kol. [15] využívají SVM regresi (*Support Vector Machines* – metoda strojového učení), zkoumající následující rysy recenzí:

- **strukturní vlastnosti** – zkoumající strukturu a formátování dokumentu, např. délku recenze, počet vět, průměrnou délku vět, procento tázacích a rozkazovacích vět a počet značek HTML pro tučný text <b> a zalomení řádku
,
- **lexikální vlastnosti** – zachycují pozorovaná slova v recenzích,
- **syntaktické vlastnosti** – procento znaků, které jsou podstatné jména, procento znaků, které jsou slovesa apod.,
- **sémantické vlastnosti** – aspekty produktu, pozitivní a negativní sentimentální slova popisující produkt nebo jeho vlastnosti,
- **vlastnosti metadat** – hodnocení recenzí (např. počet hvězdiček).

Určení kvality recenzí pomocí regresního modelu provedli také Zhang a Varadarajan [18], kteří využili podobné charakteristiky: délka recenze, hodnocení recenze, sentimentální slova, zmínění aspektů produktu apod. Na rozdíl od výše zmíněných však Liu a kol. [18] považuje za důležité 3 hlavní faktory: **odbornost recenzentů**, **včasnost recenze** a **styl recenze**. Ve své práci se autoři zaměřili na recenze filmů a pro integraci těchto faktorů navrhli nelineární

model. Ghose a Ipeirotis [6] využili tři další faktory: **profil recenzenta**, jeho **historii** (užitečnost jeho předchozích recenzí) a **čitelnost recenze** (gramatické chyby, překlepy).

Lu a kol. [20] se na tento problém zaměřili z jiného úhlu. Jejich práce je založena na dvou předpokladech: prvním z nich je skutečnost, že kvalita recenzí závisí na kvalitě recenzenta. Odhadnutí kvality autora tak může pomoci zjistit kvalitu recenze. Druhý předpoklad se týká kvality osob, s kterými je autor spojen pomocí sociálních sítí. Informace o kvalitě recenzentů mohou být získány na základě kvality jejich přátel na sociálních sítích.

Alternativní přístup Lu a kol. [20] je založen na následujících hypotézách, které zkoumají, jak se autoři recenzí chovají jednotlivě nebo uvnitř sociálních skupin:

- **hypotéza konzistentnosti autora** - recenze od toho samého autora jsou stejné nebo podobné kvality, recenzent s kvalitními recenzemi v nich pravděpodobně bude pokračovat,
- **hypotéza konzistentnosti důvěry** - vazba od recenzenta r_1 k recenzentovi r_2 je explicitní nebo implicitní prohlášení důvěry – recenzent r_1 věří recenzentovi r_2 pouze pokud kvalita recenzenta r_2 je stejná nebo vyšší než recenzenta r_1, dle [20] recenzent intuitivně nedůvěřuje někomu s nižší kvalitou recenzí, než má on sám,
- **hypotéza konzistentnosti kocitace** - tato hypotéza je založena na tom, jak jsou lidé konzistentní v důvěřování ostatním lidem - pokud dvěma recenzentům r_1 a r_2 důvěřuje třetí recenzent r_3, tak by jejich kvalita recenzentů r_1 a r_2 měla být podobná),
- **hypotéza konzistentnosti spojení** - pokud jsou dva lidé spojeni sociální sítí (r_1 věří r_2 nebo naopak), pak by jejich kvalita recenzí měla být podobná (předpokladem hypotézy je, že dva uživatelé, kteří jsou vzájemně spojeni, budou mít podobnou kvalitu recenzí spíše než dva náhodní uživatelé).

2 VYUŽITÍ ANALÝZY SENTIMENTU V OBLASTI FINANCÍ

Cílem analýzy sentimentu je prostudovat data v korpusu textových dokumentů a zjistit, zda je sentiment pozitivní nebo negativní. Index příznivosti může být měřen u oblíbenosti značky, image produktu nebo názoru na politické problémy. Zdrojem dat mohou být blogy, recenze produktů, skupinové diskuze, sociální sítě, odborné zprávy, výroční zprávy nebo jiná komunikace [34].

2.1 Zdroje sentimentu

Dle [14] mohou být zdroje sentimentu rozděleny do tří kategorií: sentiment vyjadřovaný podniky, sentiment vyjadřovaný v médiích a sentiment vyjadřovaný na internetu. V případě **sentimentu vyjadřovaném podniky** se jedná o dokumenty zveřejňované samotnými podniky, které jsou přirozeným zdrojem sentimentu. Oficiální zprávy zevnitř podniku mají určitý lingvistický styl a ráz, který může sdělit užitečné informace o očekávané budoucnosti podniku. Podniky vyjadřovaný sentiment je obzvláště užitečný při zkoumání informací o výkonnosti podniku a cen akcií. Jeho nevýhodou je však méně časté zveřejňování těchto dat, jelikož podniky obvykle publikují své zprávy jednou ročně či čtvrtletně.

Druhým typem zdroje je **sentiment vyjadřovaný v médiích**, který zahrnuje pozitiva či negativa obsažená ve zprávách, hloubkových komentářích či zprávách analytiků. Tyto informace se většinou týkají obecné ekonomické situace, celkového finančního trhu, jeho podmínek a vyhlídek, v menším měřítku individuálních firem. Informace o konkrétních podnicích mohou být využité pro analýzu cen akcií, objemu prodejů a dalších.

Posledním typem je **sentiment vyjadřovaný na internetu**. Příspěvky na internetu jsou potenciálně užitečným zdrojem sentimentu, protože mnozí lidé tráví čas na internetu čtením a psaním zpráv o akciích. Jejich zprávy mohou obsahovat užitečné postřehy, sentiment na trhu, manipulativní chování a reakce na další zdroje nových zpráv. Sentiment vyjadřovaný na internetu je označován za více zatížený šumem než předchozí typy, protože obsahuje více názorů od individuálních obchodníků. Další výhody a nevýhody jednotlivých typů shrnují následující body [14]:

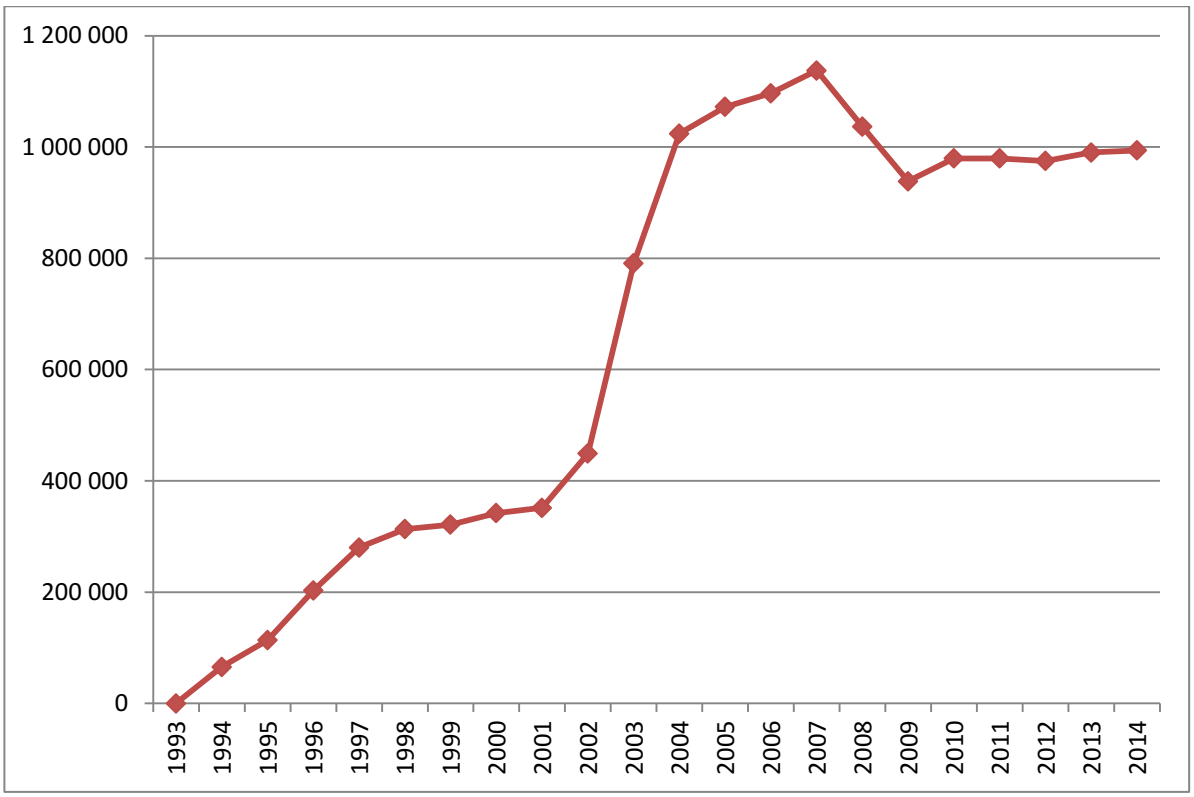
1. Dokumenty zveřejňované podniky sice obsahují informace od „zasvěcených“ pracovníků, není však jisté, že management podniku zveřejní pouze pravdu (bez snahy přilákání investorů). Na rozdíl od toho zprávy v médiích vyjadřují objektivní informace, bohužel však nejsou schopny zachytit pohled zevnitř podniku. Kvalita informací ve zprávách analytiků se nachází na pomezí výše zmíněných, jelikož

analytici mohou získat i nějaké vnitřní informace díky komunikaci s investory a dalšími účastníky trhu.

2. Zprávy v médiích pokrývají různé druhy událostí, kdežto dokumenty zveřejňované podniky obsahují pouze omezené množství informací o konkrétním podniku.
3. Zprávy v médiích zobrazují především zpětný pohled (o tom co se stalo) a ne to, co by se mohlo stát. Dokumenty podniků a zprávy analytiků mají tendenci obsahovat více výhledových prohlášení.
4. Dokumenty podniků jsou zveřejňovány obvykle ročně nebo čtvrtletně, takže sentiment z nich získaný není vhodný pro časové řady (má nízkou četnost). Je však vhodný pro zkoumání cen akcií v období zveřejňování zpráv. Články v médiích poskytují flexibilnější zdroje informací, které mohou být využívány měsíčně, týdně, denně či dokonce častěji, pokud jsou k dispozici potřebné texty.
5. Online média jsou přístupná a neregulovaná, proto mají tendenci být více zatížená šumem než klasická média a podnikové zveřejňování. Velké množství zpráv může být napsáno obchodníky zatíženými šumem nebo neinformovanými investory, kteří mohou být náchylní k určitým názorům či sentimentu a tak je jimi zveřejněná informace méně přesná nebo spolehlivá. Internetové příspěvky nejsou sice ideálním zdrojem pro informace o efektivnosti trhu, mohou však poskytnout informace o sentimentu malých investorů.
6. Předzpracování internetových příspěvků je dražší a náročnější než příprava zpráv z médií a podniků. Lidé mají na internetu tendenci psát méně formálně a jasně, takže význam jejich zpráv může být nejednoznačný. Zprávy podniků a články v médiích jsou psány profesionály a vyžadují tak méně času při předzpracování.

2.2 Analýza sentimentu ve finanční oblasti

Textová analýza ve finanční oblasti je poměrně nedávný fenomén, jehož využití se rozšířilo s rozmachem internetu a požadavky SEC na elektronické vyplňování dokumentů. SEC je zkratkou *U. S. Securities and Exchange Commission* – Komise pro cenné papíry a burzy USA. Následující Obrázek 3 zobrazuje vývoj celkového počtu dokumentů zveřejněných Komisí SEC v letech 1993 až 2014. Přesněji se jedná o počty dokumentů, které jsou veřejně dostupné online [32].



Obrázek 3: Vývoj počtu všech dokumentů zveřejněných SEC v letech 1993 - 2014

Zdroj: vlastní zpracování dle [33]

V oblasti financí byly v posledním desetiletí zkoumány dva typy sentimentu [14]. Prvním typem je sentiment investorů – tedy přesvědčení o budoucích peněžních tocích a investorském riziku, které nevychází pouze z faktů. Pomocí různých způsobů, jak změřit sentiment investora, jsou pak vyhledávány a vyčísleny efekty, jak sentiment investora ovlivňuje jednotlivé akcie a celý trh.

Druhý zkoumaný typ je sentiment založený na textu. Jedná se o sentimentální slova, vyskytující se v různých textech či dokumentech a především tedy o jejich míru pozitivnosti nebo negativnosti. Dle [14] některé studie rozlišují pouze, zda je daný výraz pozitivní či negativní. V širším měřítku však sentiment založený na textu zkoumá i intenzitu daného výrazu (silný/slabý).

Sentiment investora zachycuje subjektivní hodnocení a povahu konkrétních investorů. Toto se týká i textového sentimentu, který však zahrnuje i více objektivní odraz podmínek uvnitř firem, institucí a trhů. Spojení těchto dvou typů je složité a není přesně známo či vysvětleno, jak tyto dva typy souvisí.

Analýza sentimentu může být využita i pro zkoumání pohybu cen, je třeba se zaměřit na takové pohyby, které jsou způsobeny davovým jednáním. Je vhodné tuto analýzu

zkombinovat s technickou a fundamentální analýzou [5]. Kombinace těchto zmíněných analýz může poskytnout informace o chování trhu.

2.2.1 Ekonometrické modely, testování hypotéz

Při řešení konkrétního problému by dle [14] byl prvním krokem výběr zdroje informací (sentimentu) a rozhodnutí, zda využít přístup na základě slovníku nebo např. strojové učení (viz podkapitola 1.4). Dále je třeba vybrat model, pomocí něhož se bude zjišťovat, zda má textový sentiment dopad na lidi, instituce či trh. Rozsah možných způsobů modelování je velmi široký, některé z nich jsou uvedeny dále [14]:

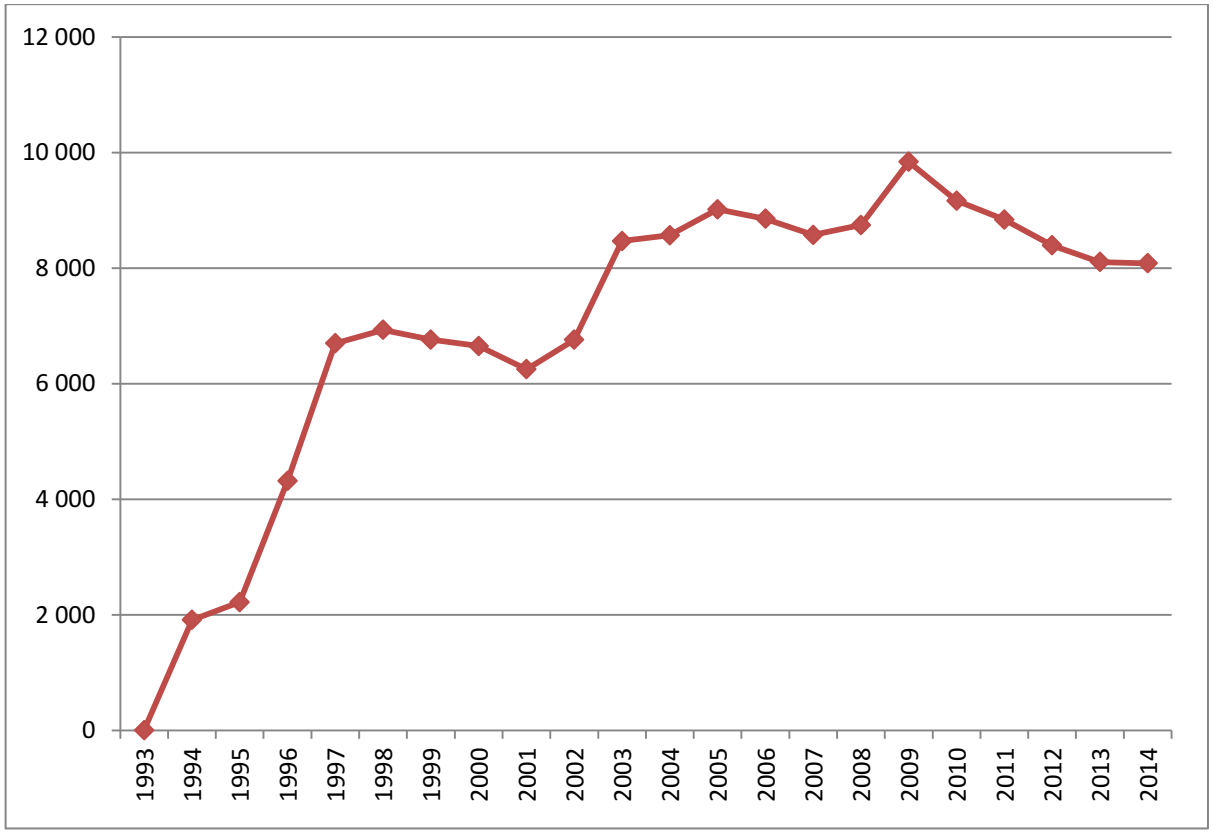
1. **Lineárně regresní model -** nejobvyklejším přístupem je aplikování lineárně regresního modelu na data časových řad, zahrnujících sentiment na všeobecné tržní úrovni a index hodnoty akcií. Model zahrnuje závislou veličinu Y, vektor nezávislých veličin $\boldsymbol{X}$ a vektor různých měřítek vlivu nebo sentimentu $\boldsymbol{S}$. Konkrétní rovnici a její vysvětlení popisuje další literatura, např. [14]. Vektor $\boldsymbol{X}$ nezávislých veličin obsahuje charakteristiky podniku a předchozí tržní proměnné jako je cash flow, poměr účetní a tržní ceny akcie, současnou a minulou hodnotu akcií či tržní hodnotu vlastního kapitálu.
2. **Vektorové autoregresní modely (VAR) -** jsou rovněž využívány pro zachycení vývoje a vzájemné závislosti mezi úrovní firem nebo trhu, nezávislými proměnnými a sentimentem. VAR jsou více pokročilé modely časových řad, všechny proměnné jsou stanoveny na základě své vlastní historie a na základě historie ostatních proměnných v modelu.
3. **Logistické a probitové regrese -** logistické (logitové) nebo probitové regrese byly použity v několika studiích pro zjištění, zda textový sentiment může pomoci předvídat nebo poznat, zda se stanou určité jevy. Tyto regrese nepředpokládají lineární závislost mezi závislými a nezávislými proměnnými a závislé proměnné musí být dichotomické (pouze 2 kategorie – např. růst nebo pokles).
4. **Modely volatility -** modely volatility mohou testovat účinky sentimentu (nebo proměnné související se sentimentem) na výnosy z akcií.
5. **Obchodní strategie na základě textového sentimentu** – je to další intuitivní způsob jak zjistit, které pozitivní a negativní výrazy vedou k významným rozdílům u výnosů z akcií. Tato metoda může být použita jako doplňková k ekonometrickým modelům, ke zjištění, jaký dopad má textový sentiment na finanční trh.

2.2.2 Vliv sentimentu na ceny a výnosy cenných papírů

Dle práce autorů Kearneyho a Liu [14] bylo zjištěno, že textový sentiment nebo ráz kvalitativních informací má dopad na ceny akcií a výnosy z nich, přičemž sentiment vyjadřovaný v médiích i na internetu má téměř okamžitý efekt. Nejsilnější vliv má negativní sentiment, jehož existence nebo velký nárůst způsobuje rychlý pokles cen.

Někteří autoři (viz [14]) zkoumáním podnikových zpráv dospěli k závěru, že tón podnikových zpráv nebo jeho změna významně koreluje s krátkodobými současnými výnosy po datu zveřejnění podnikových zpráv, např. 10-K. Dokument 10-K (*Form 10-K*) je komplexní souhrnná zpráva o výkonnosti společnosti, která musí být každý rok předložena Komisi pro cenné papíry a burzy SEC [12]. Zahrnuje více detailů než výroční zprávy, obsahuje informace o historii společnosti, vlastním kapitálu apod. Dle [14] jsou neobvyklé tržní výnosy vyšší spolu s tím, jak se tón těchto zpráv stává pozitivnější.

Vývoj počtu zpráv 10-K zveřejněných online (Obrázek 4) se v posledních letech vyvíjel obdobně jako celkový počet těchto zveřejněných dokumentů (viz Obrázek 3). Tyto zprávy tak bývají využívány mnoha autory jako zdroj sentimentu (např. [14], [19]).



Obrázek 4: Vývoj počtu dokumentů 10-K zveřejněných SEC v letech 1993 - 2014

Zdroj: vlastní zpracování dle [33]

Studie Tetlocka z roku 2007 [31] spojuje populární sloupek *Abreast of the Market* (z amerického mezinárodního deníku *Wall Street Journal*) s následným výnosem z akcií a objemem prodejů. Bylo zjištěno, že vysoký počet pesimistických slov ve sloupku předchází nižším výnosům na druhý den. Pesimismus byl určen počtem slov pomocí faktoru odvozenému ze 77 kategorií z Harvardského slovníku.

Tetlock se též zaměřil na opačný vliv - dopad změny cen či výnosů na sentiment. Spolu s Garciou se domnívají, že změny ohledně akcií se projeví v sentimentu v médiích. Na rozdíl od nich dle [14] další autoři zjistili, že hodnota akcie v daný den neovlivňuje výši sentimentu v internetových příspěvcích v den následující.

Autoři různých studií běžně používají pro klasifikaci textu externí seznamy slov jako *Harvard's General Inquirer*. Zahrnuje příklady slov pozitivních, negativních, silných, slabých, aktivních, příjemných a nepříjemných (bolestivých). Badatelé v oblasti finančnictví a účetnictví se obvykle soustředí na *GI's Harvard IV-4* kategorii negativních a pozitivních slov, ale dle [19] se zdá, že pozitivní seznam slov nemá velkou přírůstkovou hodnotu. Problémem je častá situace, kdy se podniky snaží sdělit špatnou zprávu pomocí pozitivních slov – např. místo slova „*neprospívá*" použít výraz „*nemá prospěch*". Ne vždy je pro odhadnutí tónu textu využit Harvardský slovník, ale pro klasifikaci slov se typicky využívá. Loughran a McDonald ve své práci [19] použili tento slovník, protože není chráněný a díky tomu měli přístup k tomu, která slova nejvíce přispívají k agregovaným počtům.

Loughran a McDonald poskytují ve studii *When is a Liability not a Liability?* [19] důkazy o tom, že seznam *H4N-Inf* podstatně špatně klasifikuje slova, pokud se využívá pro měření sentimentu ve finanční oblasti. Nevhodně klasifikovaná slova, např. *taxes* (daně) nebo *liabilities* (závazky), přidávají do měření šum a oslabují odhadovaný regresní koeficient. Zjistili, že téměř tři čtvrtiny (73,8 %) negativních slov podle Harvardského seznamu nejsou ve finančním kontextu typicky negativní. Slova jako například *daň*, *náklad*, *kapitál*, *závazek* nebo *cizí* jsou na Harvardském seznamu negativních slov. Stejně tak jako další slova např. *rakovina*, *pneumatika* nebo *ropa* spíše identifikují určitý segment průmyslu, než aby vyjadřovala negativní událost.

Loughran a McDonald proto vytvořili svůj vlastní seznam slov, který má typicky negativní význam ve finanční oblasti. Pro tento seznam negativních finančních slov používají název *Fig-Neg*. Některé z těchto slov se též objevují v seznamu *H4N-Inf*, ale některé ostatní ne (mezi které patří např. *felony* – zločin, *litigation* – spor, *misstatement* – mylné konstatování a *unanticipated* – neočekávaný).

Při svém výzkumu [19] vycházeli Loughran a McDonald ze zpráv 10-K z let 1994 až 2008. Vybrané podniky měli obchodovatelné akcie, jejichž cena byla alespoň 3 USD. Jedním způsobem, jak otestovat seznam slov, je zkoumat reakci trhu po zveřejnění zpráv 10-K. Pokud na tónu zprávy záleží, podniky, které vyplní 10-K velkým množstvím negativních slov, by v průměru měly pociťovat nižší výnosy.

Následující Tabulky 1 a 2 ukazují 30 nejčastějších negativních slov podle slovníku *H4N-Inf* (Tabulka 1) a podle slovníku autorů Loughrana a McDonalda (Tabulka 2). Slova označená „fajfkou" se objevují v obou seznamech negativních slov (např. *loss* – ztráta, *impairment* – zhoršení nebo *failure* – selhání).

Tabulka 1: Nejčastější negativní slova dle slovníku H4N-Inf

Panel A: H4N-Inf							
Full 10-K Document				MD&A Subsection			
Word In Fin-Neg	Word	% of Total Fin-Neg Word Count	Cumulative %	Word in Fin-Neg	Word	% of Total Fin-Neg Word Count	Cumulative %
	TAX	4.83%	4.83%		COSTS	6.45%	6.45%
	COSTS	4.61%	9.44%		EXPENSES	5.51%	11.96%
✓	LOSS	3.77%	13.21%		EXPENSE	4.70%	16.66%
	CAPITAL	3.62%	16.83%		TAX	4.68%	21.34%
	COST	3.51%	20.34%		CAPITAL	4.24%	25.58%
	EXPENSE	3.12%	23.46%		COST	3.70%	29.28%
	EXPENSES	2.92%	26.38%	✓	LOSS	3.29%	32.57%
	LIABILITIES	2.66%	29.04%		DECREASE	3.06%	35.63%
	SERVICE	2.57%	31.61%		RISK	2.97%	38.60%
	RISK	2.34%	33.95%	✓	LOSSES	2.62%	41.22%
	TAXES	2.23%	36.18%		*DECREASED*	2.21%	43.44%
✓	LOSSES	2.20%	38.38%		LIABILITIES	2.15%	45.58%
	BOARD	2.13%	40.51%		LOWER	2.10%	47.69%
	FOREIGN	1.68%	42.20%		TAXES	1.95%	49.63%
	VICE	1.52%	43.71%		SERVICE	1.91%	51.55%
	LIABILITY	1.41%	45.12%		FOREIGN	1.87%	53.42%
	DECREASE	1.29%	46.41%	✓	IMPAIRMENT	1.63%	55.05%
✓	IMPAIRMENT	1.18%	47.59%		CHARGES	1.40%	56.44%
	LIMITED	1.10%	48.69%		LIABILITY	1.16%	57.60%
	LOWER	1.01%	49.70%		CHARGE	1.16%	58.76%
✓	AGAINST	1.00%	50.70%		RISKS	1.05%	59.80%
	MATTERS	0.99%	51.69%	✓	*DECLINE*	1.00%	60.80%
✓	ADVERSE	0.94%	52.63%		DEPRECIATION	0.92%	61.72%
	CHARGES	0.94%	53.57%		MAKE	0.86%	62.58%
	MAKE	0.89%	54.46%	✓	ADVERSE	0.84%	63.42%
	ORDER	0.88%	55.33%		BOARD	0.79%	64.21%
	RISKS	0.85%	56.19%		LIMITED	0.78%	64.99%
	DEPRECIATION	0.85%	57.04%		EXCESS	0.71%	65.70%
	CHARGE	0.83%	57.87%		ORDER	0.70%	66.40%
	EXCESS	0.82%	58.69%	✓	AGAINST	0.70%	67.10%

Zdroj: [19]

Tabulka 2: Nejčastější negativní slova dle slovníku Fig-Neg

Panel B: Fin-Neg							
Full 10-K Document				MD&A Subsection			
Word in H4N-Inf	Word	% of Total Fin-Neg Word Count	Cumulative %	Word in H4N-Inf	Word	% of Total Fin-Neg Word Count	Cumulative %
✓	LOSS	9.73%	9.73%	✓	LOSS	9.51%	9.51%
✓	LOSSES	5.67%	15.40%	✓	LOSSES	7.58%	17.10%
	CLAIMS	3.15%	18.55%	✓	IMPAIRMENT	4.71%	21.81%
✓	IMPAIRMENT	3.04%	21.59%		RESTRUCTURING	2.93%	24.74%
✓	AGAINST	2.58%	24.17%	✓	DECLINE	2.89%	27.62%
✓	ADVERSE	2.44%	26.61%		CLAIMS	2.71%	30.33%
	RESTATED	2.09%	28.70%	✓	ADVERSE	2.44%	32.77%
✓	ADVERSELY	1.75%	30.45%	✓	AGAINST	2.01%	34.78%
	RESTRUCTURING	1.72%	32.17%	✓	ADVERSELY	1.94%	36.72%
	LITIGATION	1.67%	33.83%		LITIGATION	1.67%	38.40%
	DISCONTINUED	1.57%	35.40%		CRITICAL	1.63%	40.03%
	TERMINATION	1.35%	36.75%		DISCONTINUED	1.62%	41.64%
✓	DECLINE	1.19%	37.93%	✓	*DECLINED*	1.30%	42.94%
✓	CLOSING	1.08%	39.01%		TERMINATION	1.06%	44.00%
✓	FAILURE	0.97%	39.98%	✓	NEGATIVE	0.96%	44.96%
	UNABLE	0.84%	40.82%	✓	FAILURE	0.93%	45.89%
✓	*DAMAGES*	0.82%	41.64%		UNABLE	0.91%	46.80%
✓	DOUBTFUL	0.77%	42.41%	✓	CLOSING	0.86%	47.65%
✓	LIMITATIONS	0.75%	43.17%		*NONPERFORMING*	0.81%	48.47%
✓	FORCE	0.74%	43.91%	✓	IMPAIRED	0.81%	49.28%
✓	VOLATILITY	0.73%	44.64%	✓	VOLATILITY	0.79%	50.07%
	CRITICAL	0.73%	45.37%	✓	FORCE	0.75%	50.82%
✓	IMPAIRED	0.70%	46.07%	✓	*NEGATIVELY*	0.73%	51.56%
	TERMINATED	0.70%	46.77%	✓	DOUBTFUL	0.72%	52.27%
✓	*COMPLAINT*	0.63%	47.39%	✓	*CLOSED*	0.70%	52.97%
✓	DEFAULT	0.57%	47.96%	✓	DIFFICULT	0.69%	53.66%
✓	NEGATIVE	0.51%	48.47%	✓	*DECLINES*	0.63%	54.29%
✓	*DEFENDANTS*	0.51%	48.99%	✓	*EXPOSED*	0.60%	54.89%
✓	*PLAINTIFFS*	0.51%	49.49%	✓	DEFAULT	0.59%	55.48%
✓	DIFFICULT	0.50%	50.00%	✓	*DELAYS*	0.56%	56.04%

Zdroj: [19]

V levé části obou tabulek se nacházejí nejčastější negativní slova z celé zprávy 10-K, pravá část zobrazuje slova pouze z části MD&A. MD&A (*Management Discussion and Analysis*) je část výroční zprávy podniku 10-K, ve které management rozebírá různé aspekty podniku – jak minulé, tak budoucí [13]. Jak lze vidět na obou obrázcích, pořadí negativních slov se liší v závislosti s tím, zda je zdrojem sentimentu celá zpráva 10-K nebo pouze jedna její položka.

Výsledkem studie Loughrana a McDonalda [19] bylo, že medián výnosů u seznamu *H4N-Inf* neodrážel konzistentní vztah s podílem negativních slov. Firmy s vysokým podílem negativních slov (dle Harvardského seznamu) mají pouze o něco nižší návratnost než firmy s relativně nižším počtem negativních slov. Na rozdíl od toho firmy zkoumané dle seznamu *Fig-Neg* vykazovaly klesající výnosy s tím, jak se zvyšoval počet negativních slov.

Loughran a McDonald tak dokázali, že využití obecného seznamu pro finanční účely může vést k nesprávné klasifikaci a špatné korelaci. Veškerá textová analýza dle nich stojí či padá na procesu kategorizace.

2.2.3 Vliv sentimentu na ostatní tržní proměnné

Některé studie rovněž prokázaly, že textový sentiment má výrazný dopad na objem obchodů. Tetlock [31] zjistil, že neobvykle vysoká nebo nízká hladina pesimismu dočasně vede k velkému množství tržních obchodů. Další autoři [14] objevili, že internetové příspěvky mohou předpovídat objem obchodů a pomocí denních dat ukazují, že vliv internetových příspěvků na objem obchodů má větší dopad než vliv objemu obchodů na internetové příspěvky. Některými autory byl dokázán vztah mezi sentimentem a objemem obchodů. I když tito autoři zkoumali internetové příspěvky, usuzují, že nelze předvídat konkrétní vliv sentimentu na budoucí objem obchodů [14].

Další studie objevily důkazy o vztahu mezi textovým sentimentem a volatilitou trhu s akciemi. Volatilita může být chápána jako míra kolísání určité hodnoty. Čím vyšší je volatilita, tím vyšší je riziko a tím hůře se předpovídá budoucí výnos [27]. Dle [14] někteří autoři zjistili, že internetové příspěvky mohou předpovídat volatilitu. Dále bylo zjištěno, že pozitivní i negativní novinky v médiích mají dopad na volatilitu výnosů firem a když management předloží negativní zprávy, volatilita výnosů roste.

2.2.4 Vliv sentimentu na fundamentální ukazatele a efektivnost trhu

Průzkumy provedené v současné době dokazují, že textový sentiment koreluje s budoucími charakteristikami podniku a jeho výkonem. Studie Tetlocka a kol. [30] ukazuje, že negativní slova v novinových zprávách předpovídají nízké firemní zisky. Stejně tak vyšší míra pesimismu v sekci MD&A je spojena s nižší budoucí hodnotou rentability aktiv [14].

2.2.5 Predikce finančních podvodů a bankrotů pomocí analýzy sentimentu

Cecchini a kol. [2] aplikovali vlastní metodologii na dvě důležité finanční události – finanční podvod a bankrot. Vytvořili vlastní slovníky klíčových slov, které by pomohli předpovídat podvody a bankroty na základě textu MD&A ze zpráv 10-K. Tato metoda by tak mohla sloužit vládním agenturám pro zjišťování podvodů, či investorům pro odhad rizika bankrotu jednotlivých firem.

Prvním krokem je podle [2] analýza textu (ať už manuální či automatická). Někdy též může být stanovena hypotéza o struktuře textu. Dále je zapotřebí vytvořit první korpus, sebráním dokumentů od firem, které jsou předmětem zájmu (například firmy v bankrotu), a druhý korpus firem, kterých se tento problém netýká. Automatická procedura analyzuje obsah textu a vytvoří slovník klíčových slov, který rozliší dva druhy dokumentů. Také se zjišťuje počet klíčových slov, s cílem vytvořit funkci pro klasifikaci dvou tříd. Tato funkce je testována na sadě dokumentů, protože přesnost klasifikační funkce pomáhá určit hodnotu slovníku.

Dle Cecchiniho a kol. [2] je přínosem jejich práce vývoj metodologie zaměřené speciálně na predikci finančních událostí. Na začátku byla vytvořena textová metodologie a její účinnost byla testována na dvou odlišných finančních událostech. Nástrojem byl program vyvinutý pomocí výpočetních lingvistických teorií bez lidského zásahu. Bylo dokázáno, že textové informace ve výročních zprávách mohou být využity pro účely detekce finančních událostí. Srovnávacím měřítkem bylo porovnání výsledků textové metody s tradičními predikčními metodami (s využitím kvantitativních finančních proměnných). Výsledky dokázaly, že textové informace jsou konkurenceschopné kvantitativním metodám. Též bylo zjištěno, že nejlepších výsledků lze dosáhnout kombinací obou metod, mezi textovými a kvantitativními informacemi existuje doplňkový vztah.

Pro otestování metodologie shromáždili autoři dvě sady dat pro zmíněné důležité finanční události – bankrot a finanční podvod. Ověření v těchto dvou odlišných situacích pomáhá dokázat potenciál metodologie. Bankrot a finanční podvod byly dle autorů vybrány, protože obě situace mají katastrofální dopad na hodnotu výnosu akcionáře. Základními daty pro bankrot bylo 78 podniků s bankrotem v letech 1994 až 1999 a stejně tak 78 kontrolních podniků bez bankrotu. Pro finanční podvod bylo vybráno 61 podniků s finančním podvodem a 61 podniků bez něj, z let 1993 až 2002.

Získaná data byla statisticky zpracována. U podniků s bankrotem i bez něj (stejně tak pro podniky s finančním podvodem a bez něj), byly zjišťovány následující ukazatele: průměr (*mean*), směrodatná odchylka (*st. dev.)* a *p*-hodnota (*p-value*) celkových aktiv (*total assets*), tržeb (*sales*), tržní hodnoty (*market value*), běžné likvidity (*current ratio*) a rentability aktiv (*return on assets*).

Pro podniky s bankrotem a bez něj, byla při hladině významnosti α=0,05 významně odlišná pouze rentabilita aktiv a tržní hodnota. Tržní hodnota je nižší u podniků s horším finančním zdravím (jako u těch, u kterých nastane během jednoho ruku bankrot). Rentabilita

aktiv u podniků s bankrotem má tendenci být nižší, protože podniky využívají aktiva neefektivně nebo vyrovnávají ztrátu.

U podniků s finančním podvodem a bez něj ukazuje *p*-hodnota pro celková aktiva, tržby a tržní hodnotu, že vzorky nejsou významně odlišné. Rentabilita aktiv byla výrazně nižší u podniků s podvodem. Běžná likvidita byla na skoro hranici s hodnotou 0,069. Tyto výsledky byly očekávány, protože podniky mají tendenci k podvodu, pokud jsou pod zvyšujícím se finančním tlakem.

U obou druhů problémových podniků (těch v bankrotu a s podvodem) byl vytvořen seznam klíčových slov, odlišujících firmy s problémem a bez něj. *Čistý příjem*, *hrubá marže* a *výzkum a vývoj* jsou první mezi seznamem slov pro podezření bankrotu. Celkem 7 prvních spojení v seznamu odkazuje na finanční proměnné. Pro druhou část podniků (s finančním podvodem) je diskriminátor s největší váhou *konec roku prosinec* a za ním *konec roku*. Na třetím místě se umístilo spojení *společnost má* a na čtvrtém místě *výdaje na výzkum a vývoj*.

V konečném výsledku bylo zjištěno, že slovníky vytvořené Cecchinim a kol. jsou schopny určit podniky s finančním podvodem v 75 % případů a firmy s bankrotem v 80 % případů [2]. Tyto výsledky byly porovnány s výsledky kvantitativních predikčních metod (*Beneishův model* pro finanční podvod a *Altmanův model* pro bankrot). Při využití stejných dat dosáhla lepších výsledků metodologie vytvořená Cecchinim a kol. Dále byly tyto metody vzájemným sloučením kvantitativních dat a textu spojeny a testovány, zda se vzájemně doplňují. Tímto spojením bylo dosaženo lepších výsledků v obou případech – u bankrotu bylo dosaženo přesnosti 83,87 % a u podvodu 81,97 %. Text v MD&A sekci tedy obsahuje informace, které jsou doplňkové pro kvantitativní data. Dle [2] může být tato metodologie aplikována na dostupné texty i pro ostatní finanční problémy.

Spojení kvantitativních finančních ukazatelů a analýzy sentimentu se ukázalo přesnější také při predikci Altmanova Z-skóre [8] a predikci úvěrového ratingu [7]. Cílem Altmanova modelu finančního zdraví (Altmanova Z-skóre) je zjistit, zda je podnik ohrožen bankrotem. Model je založen na poměrových ukazatelích, které se využívají např. ve finanční analýze. Na základě hodnoty Z-skóre je podnik zařazen do jedné ze tří skupin („bezpečná zóna“, „šedá zóna“, „krizová zóna“). Pouze na základě tohoto výsledku však není možné pochopit všechny okolnosti, které k němu vedly. Dle autorů [8] není podstatné, zda bude pro hodnocení podniku vybrán Altmanův či jiný model, důležitý je vliv sentimentu na vnímání společnosti zúčastněnými stranami.

Na predikci úvěrového ratingu se zaměřili ve své práci Hájek a Olej [7]. Rating zkoumaných podniků byl určen podle ratingové agentury *Standard & Poor's* (podniky spadaly buď pod investiční, nebo neinvestiční stupeň). Pomocí metod neuronových sítí a metod strojového učení byla nejprve zkoumána přesnost predikce úvěrového ratingu pomocí finančních ukazatelů. Poté byla zjišťována přesnost těchto ukazatelů ve spojení s analýzou sentimentu. Díky analýze sentimentu bylo u všech metod dosaženo vyšší přesnosti (toto zlepšení bylo ve většině případů statisticky významné).

3 SBĚR A PŘÍPRAVA DAT

V této diplomové práci byly analyzovány konkrétní podniky pro otestování vlivu sentimentu na změnu ceny akcií. Pro tento účel byly vybrány americké banky, splňující kritérium velikosti celkové tržní hodnoty alespoň 100 milionů amerických dolarů. Toto kritérium splnilo zhruba 180 bank, pro které bylo třeba dále zajistit jejich nejnovější výroční zprávy – formuláře 10-K. Pouze u několika málo bank byly k dispozici údaje za rok 2014, nakonec proto byly vybrány zprávy za rok 2013, aby byly všechny získané údaje o bankách porovnatelné.

3.1 Hodnocení sentimentu

U všech bank nebylo možné výroční zprávy 10-K za rok 2013 dohledat, jejich počet se tak snížil na 132. Výroční zprávy bank byly stahovány z internetových stránek Komise pro cenné papíry a burzy USA (*U.S. Securities and Exchange Commission*) [32]. Z celé výroční zprávy 10-K bylo pro analýzu sentimentu třeba vybrat pouze její specifickou část, aby výsledky byly co nejméně zkreslené. Konkrétně se jednalo o položku 7 (*Item 7)* - *Management Discussion & Analysis* (MD&A). Jak již bylo zmíněno v předchozím textu této práce, MD&A je ta část zpráv 10-K, kde management diskutuje o různých, minulých i budoucích, aspektech podniku. Položka 7 poskytuje shrnutí informací o operacích v předchozím roce a o tom, jak se během této doby podniku dařilo [13]. Management se v této části věnuje rovněž nadcházejícímu roku – navrhuje budoucí cíle a přístupy k novým projektům. Tato část obsahuje užitečné informace, avšak není auditovaná.

Získané položky 7 byly poté pomocí různých slovníků slov zpracovány v softwaru STATISTICA. Použity byly slovníky od různých autorů. Prvním variantou byly seznamy pozitivních a negativních slov od Elaine Henry [10], druhou možností seznamy pozitivních, negativních, neurčitých, právnických, silných a slabých modálních slov od autorů Loughrana a McDonalda [19]. Oba tyto slovníky byly navrženy speciálně pro finanční oblast.

Pomocí karty *Data mining* a funkce *Text Mining* lze v programu STATISTICA snadno zjistit počty slov ze slovníku. Prvním krokem je výběr zdrojového textu (položek MD&A u celkem 132 bank) a dále slovníku, podle kterého bude STATISTICA slova v dokumentech vyhledávat. Výsledkem je výčet kořenů slov spolu s jejich absolutními počty. Kromě kořenů uvádí STATISTICA i příklady konkrétních slov, která jsou z daného kořenu odvozena.

Absolutní počty slov však nejsou pro další zpracování vyhovující, hlavně vzhledem k rozdílné délce jednotlivých dokumentů. Proto je třeba využít relativních (vážených) četností. Stanovení vah slov může mít velký dopad na konečné výsledky. Systémy pro stanovení vah většinou řeší tři komponenty [19]:

- četnost slova vzhledem k dokumentu (která může být často měřena poměrným výskytem nebo záznamem četnosti),
- formu normalizace pro délku dokumentu,
- četnost slova vzhledem k celému korpusu (typicky měřené inverzní frekvencí výskytu slova v dokumentu).

Váhové systémy se obvykle označují výrazem *tf.idf* nebo (*TF-IDF, Term Frequency Inverse Document Frequency*). První část výrazu (*tf*) označuje četnost slova, druhá část (*idf*) inverzní frekvenci výskytu slova v dokumentu. Loughran a McDonald použili ve své práci [19] jeden z nejčastěji používaných systémů, s modifikací, která se uzpůsobuje délce dokumentu:

$$w_{i,j} = \begin{cases} \frac{\left(1 + log\left(tf_{i,j}\right)\right)}{\left(1 + log(a)\right)} log \frac{N}{df_i} & \text{pro } tf_{i,j} \geq 1, \\ 0 & \text{jinak} \end{cases} \tag{7}$$

kde N představuje celkový počet dokumentů, df_i počet dokumentů obsahujících alespoň jeden výskyt i-tého slova, $tf_{i,j}$ absolutní počet i-tého slova v j-tém dokumentu a a představuje průměrný počet slov v dokumentu.

Pro účely této práce bylo vážení pojmů provedeno přímo v programu STATISTICA. Místo absolutního (*raw*) počtu, byla při zpracování vybrána funkce inverzní frekvence výskytu v dokumentu (*inverse document frequency*). Vzorec pro tuto funkci je následující:

$$idf_i = log \frac{N}{df_i} \tag{8}$$

kde N je celkový počet dokumentů a df_i je počet dokumentů obsahující alespoň jeden výskyt i-tého slova.

3.1.1 Slovníky Elaine Henry

Elaine Henry ve své práci *Are Investors Influenced by the Way Earnings Press Releases are Written?* [10] uvádí dva seznamy slov – pozitivních a negativních – jednoduše přímo v textu vypsaných. Oba seznamy obsahují shodně 85 slov, z nichž některá mají shodný kořen slova. Jako příklad pozitivních slov lze uvést *pozitivní*, *pozitiva*, *úspěch*, *úspěšný*, *silný*, *síla* apod. Obdobně mezi negativní slova patří *negativní*, *negativa*, *selhat*, *selhání*, *pokles*, *klesající* atd. Seznam všech pozitivních i negativních slov stanovených podle Elaine Henry zobrazuje následující Obrázek 5.

POSITIVITY word list
positive positives success successes successful succeed succeeds succeeding succeeded accomplish accomplishes accomplishing accomplished accomplishment accomplishments strong strength strengths certain certainty definite solid excellent good leading achieve achieves achieved achieving achievement achievements progress progressing deliver delivers delivered delivering leader leading pleased reward rewards rewarding rewarded opportunity opportunities enjoy enjoys enjoying enjoyed encouraged encouraging up increase increases increasing increased rise rises rising rose risen improve improves improving improved improvement improvements strengthen strengthens strengthening strengthened stronger strongest better best more most above record high higher highest greater greatest larger largest grow grows growing grew grown growth expand expands expanding expanded expansion exceed exceeds exceeded exceeding beat beats beating

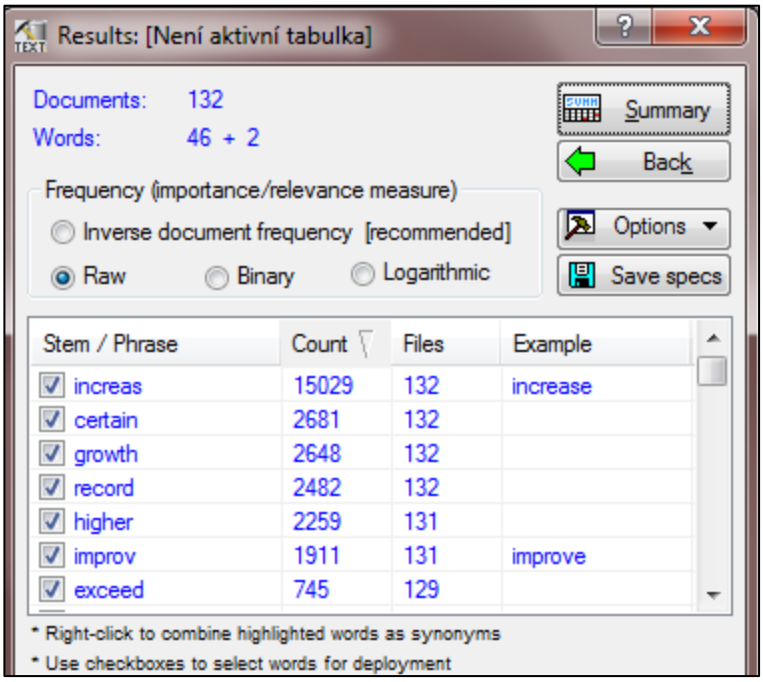
NEGATIVITY word list
negative negatives fail fails failing failure weak weakness weaknesses difficult difficulty hurdle hurdles obstacle obstacles slump slumps slumping slumped uncertain uncertainty unsettled unfavorable downturn depressed disappoint disappoints disappointing disappointed disappointment risk risks risky threat threats penalty penalties down decrease decreases decreasing decreased decline declines declining declined fall falls falling fell fallen drop drops dropping dropped deteriorate deteriorates deteriorating deteriorated worsen worsens worsening weaken weakens weakening weakened worse worst low lower lowest less least smaller smallest shrink shrinks shrinking shrunk below under challenge challenges challenging challenged

Obrázek 5: Pozitivní a negativní slova dle Elaine Henry

Zdroj: [10]

Při aplikaci pozitivního seznamu slov na texty bank bylo celkem objeveno 46 pozitivních slov. Obrázek 6 ukazuje kořeny slov, seřazené podle četnosti jejich výskytu v dokumentech. Nejčastěji se v souborech vyskytoval kořen slova *increas* (celkem 15 029 výskytů ve všech souborech) – konkrétně se jednalo o slova *increase* (zvýšit), *increased* (zvýšený), *increases* (zvyšuje) a *increasing* (zvyšující se). Dalším častým výrazem byl kořen *certain* (jistý), který se v textech vyskytl již podstatně méně (celkem 2 681krát). Pouze dvakrát se v textech

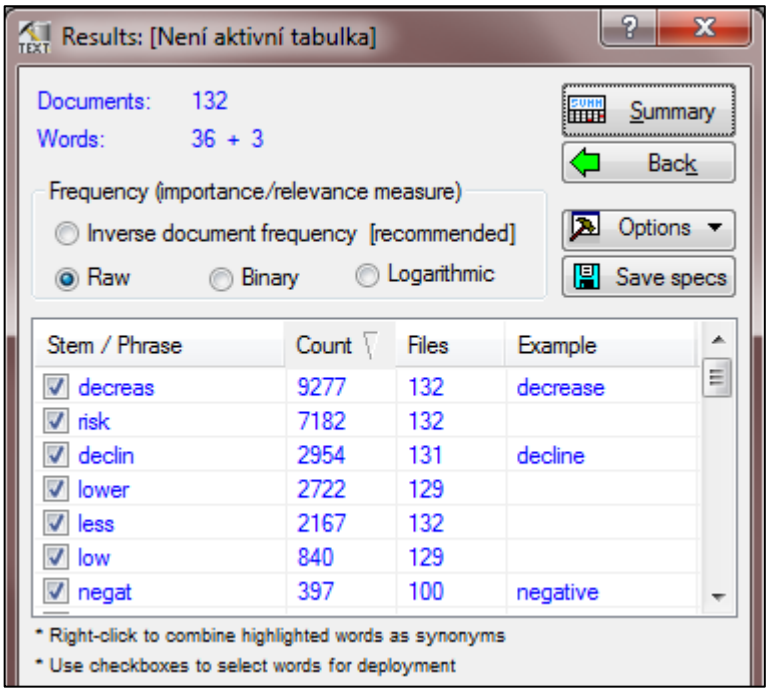
vyskytl výraz *leader* (vůdce). Při využití funkce *inverse document frequency* byl nejvýznamnějším výrazem *grew* (rostl).



Obrázek 6: Počet pozitivních slov dle slovníku Elaine Henry

Zdroj: vlastní zpracování v programu STATISTICA

Následně byly stejným způsobem zkoumány i počty negativních slov. Celkem bylo objeveno 36 kořenů slov, např. jak ukazuje Obrázek 7, kořen *declin*, který se objevil téměř ve všech souborech kromě jednoho a to celkem 2 954krát. Nejčastějším negativem byl kořen *decreas* včetně slov *decrease* (pokles), *decreased* (poklesl), *decreases* (snižuje) a *decreasing* (klesající). Druhým nejčastějším výrazem bylo slovo risk s počtem 7 182 výskytů. Nejméně se v textech objevil výraz *disappoint* (zklamat) a to pouze dvakrát. S využitím inverzní četnosti výskytu slov v dokumentu byl významným kořen *deterior* se slovy *deteriorate* (zhoršit), *deteriorates* (zhoršuje), *deteriorating* (zhoršující), *deteriorated* (zhoršil).



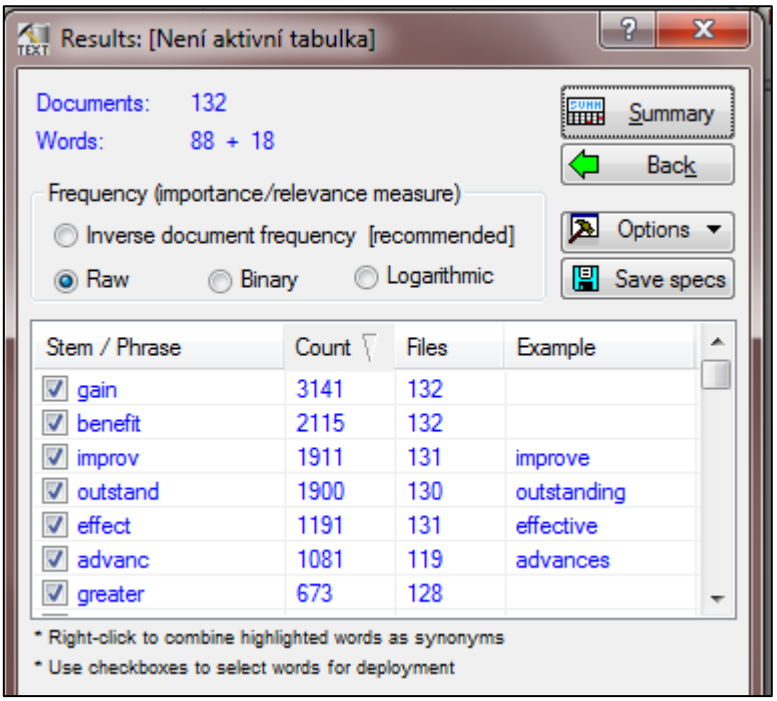
Obrázek 7: Počet negativních slov dle slovníku Elaine Henry

Zdroj: vlastní zpracování v programu STATISTICA

3.1.2 Slovníky Loughrana a McDonalda

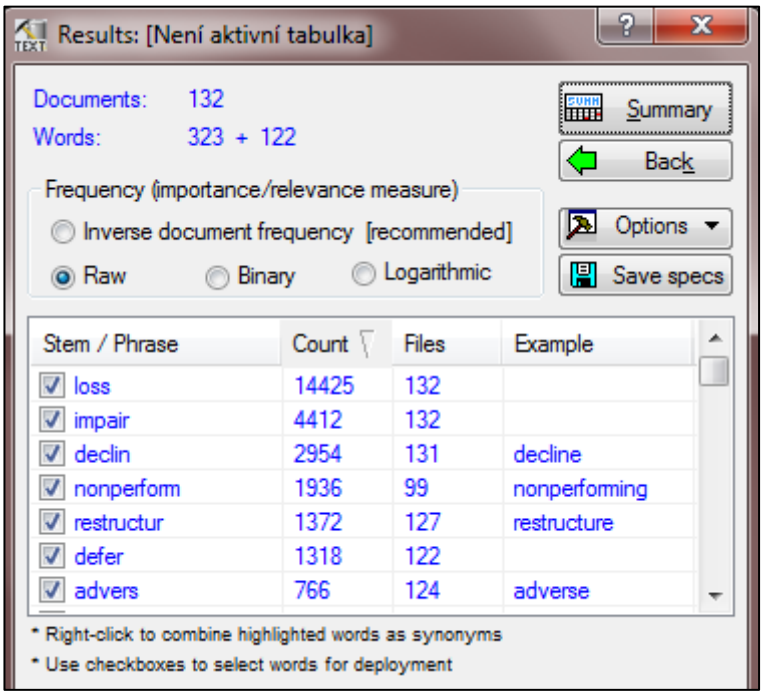
Slovníky vytvořené Loughranem a McDonaldem byly využity v jejich práci *When is a Liability not a Liability?* [19], která již byla zmíněná několikrát v předchozích kapitolách této práce. Autoři vytvořili seznamy několika druhů slov, ne pouze pozitivních a negativních, ale též neurčitých (*uncertainty*), právnických (*litigious*), modálních silných slov (*modal strong*) a modálních slabých slov (*modal weak*). Většina z těchto kategorií obsahuje více slov než slovníky dle Elaine Henry. V kategorii neurčitých nachází 291 slov, pozitivních 357, právnických 871 a seznam negativních slov zahrnuje dokonce 2 349 slov. Seznamy modálně silných a slabých slov obsahují shodně 19 slov.

Nejprve byl zkoumán v textech MD&A bank počet pozitivních slov. Celkem bylo nalezeno 88 kořenů pozitivních slov, některé z nich ukazuje Obrázek 8. S počtem výskytů 3 141 byl nejpočetnějším kořenem výraz *gain* (získat), spolu se slovy *gained* (získané), *gaining* (získávání) a *gains* (zisky). Dalším častým slovem byl výraz *benefit* (výhoda) a slova tvořená z tohoto kořenu. Naopak slova *win* (výhra, vítězství), *distinct* (zřetelný) nebo *bolster* (podpořit) se ve všech dokumentech vyskytovaly vždy pouze dvakrát. Po využití inverzní funkce bylo významným výrazem slovo *progress* (pokrok).



Obrázek 8: Počet pozitivních slov dle slovníku Loughrana a McDonalda

Zdroj: vlastní zpracování v programu STATISTICA



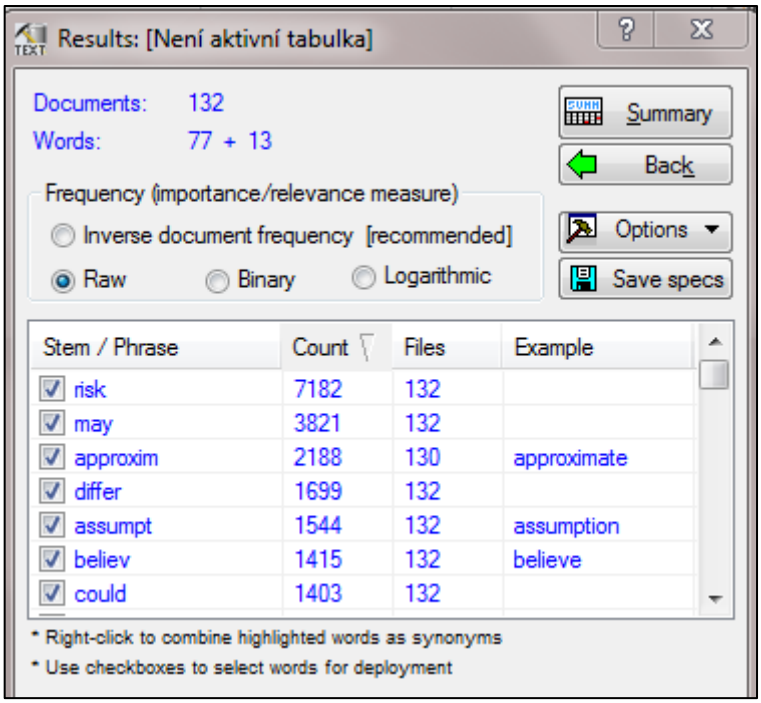
Obrázek 9: Počet negativních slov dle slovníku Loughrana a McDonalda

Zdroj: vlastní zpracování v programu STATISTICA

Jako další přišly na řadu opět negativní výrazy. Dle Obrázku 9 (viz výše) bylo nalezeno 323 negativních kořenů. S přehledem nejčastějším negativním kořenem byl v tomto případě výraz *loss* (ztráta), spolu s odvozeným *losses* (ztráty). Tato slova se objevila ve všech 132 dokumentech a to dokonce 14 425krát. Druhý nejčastější výraz byl kořen *impair*

(znehodnotit, zhoršit), který se v textu objevil 4 412krát. S využitím *inverse document frequency* bylo nejdůležitějším výrazem slovo *unfunded* (nefinancovaný).

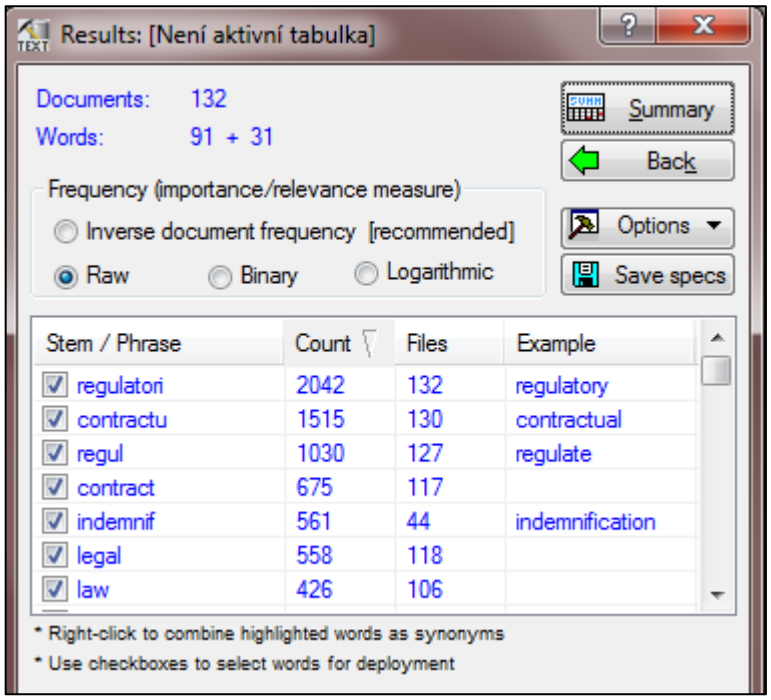
Poté bylo na řadě prozkoumat další kategorie slov – např. neurčité výrazy, kterých software STATISTICA objevil celkem 77 (viz Obrázek 10). Nejčastějším neurčitým výrazem bylo slovo *risk* (riziko) spolu s *risked* (riskoval), s výskytem 7 128krát. Dalším častým výrazem bylo *may* (možná, třeba). Jeden z méně častých výrazů byl kořen *ambigu* se slovem *ambiguous* (dvojznačný), který se objevil pouze ve dvou dokumentech, v každém pouze jednou.



Obrázek 10: Počet neurčitých slov dle slovníku Loughrana a McDonalda

Zdroj: vlastní zpracování v programu STATISTICA

Další početnou kategorií byla slova právnická, u kterých bylo objeveno 91 kořenů. Jak ukazuje Obrázek 11, nejpočetnějším kořenem byl výraz *regulatory*, tedy regulační, který se objevil v každém z dokumentů alespoň jednou, celkem 2 042krát. Pouze zhruba o čtvrtinu méně častým výskytem (1 515krát) byl výraz *contractu* se slovy *contractual* (smluvní) a *contractually* (smluvně), které se objevily ve 130 dokumentech. Ze všech 91 právnických výrazů se 13 z nich objevilo v dokumentech pouze 2krát – např. *therefrom* (z toho). Funkce *inverse document frequency* označila za nejvýznamnější právnický výraz kořen *indemnif* s výrazy *indemnify* (odškodnit) a *indemnifying* (odškodnění).

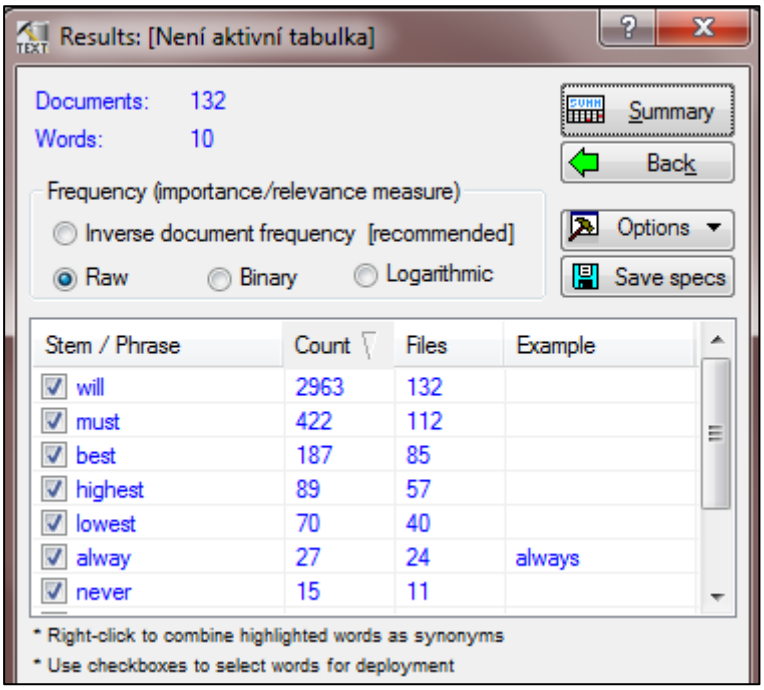


Obrázek 11: Počet právnických slov dle slovníku Loughrana a McDonalda

Zdroj: vlastní zpracování v programu STATISTICA

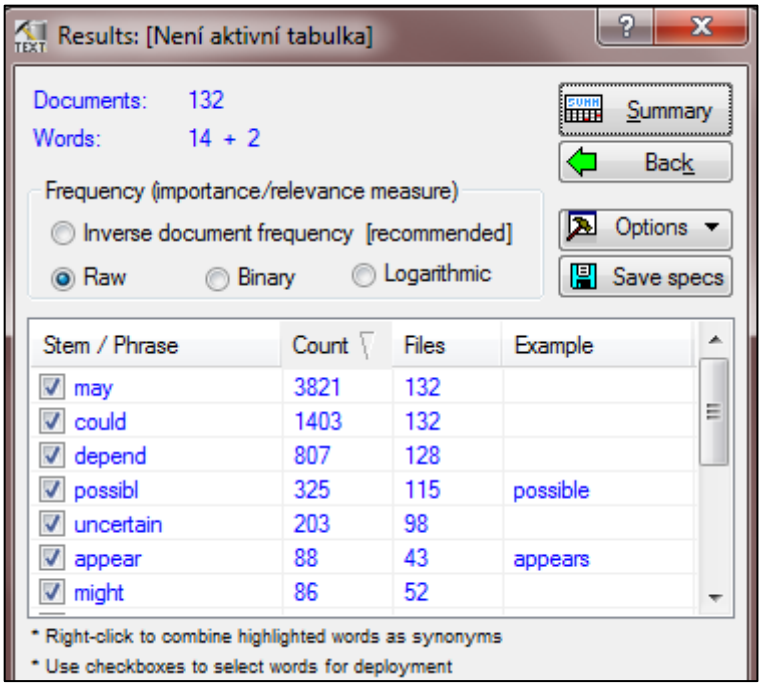
Posledními kategoriemi jsou méně početná modální silná a slabá slova. Obrázek 12 ukazuje většinu z modálních silných slov, kterých bylo v textu nalezeno pouze 10. Jednoznačně nejvýznamnějším výrazem byl výraz *will*, vyjadřující budoucí čas. Objevoval se všech dokumentech celkem 2 963krát. S využitím inverzní funkce bylo nejvýznamnějším slovem slovo *lowest* (nejnižší).

O něco více, celkem 14, bylo objeveno modálně slabých slov (viz Obrázek 13). Z nich se nejčastěji objevovalo slovo *may* (moci, smět), celkem 3 821krát. Druhé místo obsadil výraz *could* (mohl, mohl by apod.) s počtem 1 403. Nejméně častým výrazem bylo slovo *perhaps* (možná), které podniky využili 7krát. Funkcí *inverse document frequency* bylo zjištěno, že nejdůležitějším modálně slabým slovem je slovo *appear* (zdá se).



Obrázek 12: Počet silných modálních slov dle slovníku Loughrana a McDonalda

Zdroj: vlastní zpracování v programu STATISTICA



Obrázek 13: Počet slabých modálních slov dle slovníku Loughrana a McDonalda

Zdroj: vlastní zpracování v programu STATISTICA

S využitím funkce inverzní četnosti výskytu slov v dokumentu se za každou banku provedl součet všech hodnot sentimentu a následně se vydělil počtem zjištěných sentimentálních slov. Takto byla zjištěna průměrná hodnota sentimentu jednotlivého dokumentu. Stejný postup se opakoval u všech slovníků a seznamů slov.

3.2 Ostatní ukazatele

K sentimentálním veličinám z předchozí podkapitoly byly ve shodě s předchozími studiemi (např. [10], [19]) přidány další ukazatele – tržní kapitalizace, objem akcií, typ burzy, P/E, P/B, ROE a poměr celkového dluhu k celkovým aktivům. Hodnoty ukazatelů byly získány z [21]. Všechny údaje se vztahují k roku 2013, stejně jako výroční zprávy bank.

Prvním z přidaných ukazatelů je **tržní kapitalizace**, která vyjadřuje násobek tržní ceny akcie a počtu emitovaných akcií (viz vzorec 9) [24]. Hodnoty tržní kapitalizace u všech bank jsou uvedeny v milionech USD.

$$tržní\ kapitalizace = \ tržní\ cena\ akcie * počet\ emitovaných\ akcií \tag{9}$$

Druhým ukazatelem je **objem akcií** obchodovaných na burze. Pro každý podnik byla zjišťována velikost objemu akcií pět dní před zveřejněním jejich závěrečné zprávy. Dalším ukazatelem byl zvolen **typ burzy**. Pro burzu New York Stock Exchange (NYSE) byla zvolena hodnota 0, pro burzovní trh NASDAQ hodnota 1.

Čtvrtým ukazatelem je hodnota **ukazatele P/E**:

$$P/E = \frac{tržní\ cena\ akcie}{zisk\ na\ akcii} \tag{10}$$

P/E, zkratka pro *Price-Earnings Ratio*, vyjadřuje poměr aktuální tržní ceny akcie a zisku na tuto akcii [16]. Když je hodnota P/E podniku výrazně nižší než průměr v odvětví, akcie podniku mohou být podhodnoceny. Naopak pokud je P/E vyšší než průměr v odvětví, akcie mohou být nadhodnoceny. Ukazatel P/E bývá nedílnou součástí burzovních zpráv, promítá se v něm budoucí očekávání investorů, týkající se tempa růstu, míry zisku nebo podílu dividend na zisku. Všeobecně přijatelná hodnota ukazatele se pohybuje v rozmezí 8 až 12 [26], vyjadřuje, za kolik let se akcie sama zaplatí. U některých perspektivních nebo atraktivních akcií se může objevovat hodnota 15 či vyšší. Dlouhodobě jsou však tyto výsledky neudržitelné a mohou signalizovat budoucí pokles kurzu. Ukazatel P/E je mezipodnikově srovnatelný, je však třeba brát v úvahu odlišnosti průměrných hodnot v jednotlivých odvětvích.

Výslednou hodnotu P/E mohou zkreslit některé faktory [26]:

- **použité účetní metody** – velikost zisku podniku závisí i na použitých účetních metodách, které se mohou lišit v jednotlivých zemích,
- **jednorázové obchodní a finanční operace** – především prodej části podniku, mimořádné odpisy apod.,

- **aktuálnost dosazovaných údajů** – u tržní ceny akcie se mohou vyskytnout pochybnosti pouze u veřejně neobchodovatelných akcií nebo u akcií s velmi nízkou likviditou, větší problém je s aktuálností čistého zisku na akcii - největší riziko nastává na začátku účetního období (před účetní závěrkou), kdy podniky využívají zastaralé údaje z předminulého roku.

Podobně **ukazatel P/B** (nebo P/BV), *Price to Book Value*, porovnává tržní cenu akcie a účetní hodnotu vlastního kapitálu na akcii (viz vzorec 11) [16].

$$P/B = \frac{tržní\ cena\ akcie}{účetní\ hodnota\ vlastního\ kapitálu\ na\ akcii} \tag{11}$$

Hodnota ukazatele P/B nižší než 1 signalizuje, že perspektivy podniku nemusí být dobré, akcie podniku mohou být podhodnoceny. V případě, kdy je akcie podhodnocena, je výhodné tuto akcii nakoupit. Pokud je hodnota ukazatele P/B vyšší než 1, investoři si v daném okamžiku cení více akcie než podílu hodnoty majetku podniku [26]. Tržní cena akcie odráží aktuální názor investorů, kdežto účetní hodnoty jsou ovlivněny účetními metodami podniků, použitými při oceňování aktiv, a nemusí tak být objektivní. U tržní ceny akcie se však doporučuje si momentální kurz prověřit, zda je dlouhodobě stabilní či volatilní. Momentální extrémně nízké nebo naopak vysoké hodnoty kurzu mohou hodnotu P/B zkreslovat. Další nevýhodou ukazatele P/B je, že neinformuje o budoucích výnosových perspektivách, k nimž přihlížejí např. dlouhodobí investoři. Hodnoty ukazatele P/B také nejsou meziodvětvově či mezioborově porovnatelné.

Ukazatel ROE (*Return on Equity*) – rentabilita vlastního kapitálu - je šestým a předposledním ukazatelem. Rentabilita vlastního kapitálu je vypočtena poměrem čistého zisku (EAT) a vlastního kapitálu:

$$ROE = \frac{EAT}{vlastní\ kapitál} \tag{12}$$

ROE tak vyjadřuje výnosnost kapitálu vloženého vlastníky podniku [16]. Hodnota ukazatele by neměla být nižší, než je výnosnost státních dluhopisů – tedy než je úroveň bezrizikové úrokové míry existující na finančním trhu [26]. Rentabilita vlastního kapitálu by měla být vyšší než rentabilita celkových aktiv (ROE > ROA).

Posledním ukazatelem je **poměr celkového dluhu a celkových aktiv podniku**, který byl u většiny bank získán z [21], v několika málo případech byl vypočten z příslušných ukazatelů v rozvaze.

4 PREDIKCE ZMĚNY CENY AKCIE POMOCÍ ANALÝZY SENTIMENTU

4.1 Regresní analýza

Pro zjištění vlivu sentimentu na změnu cen akcií je v této práci využita jedna z nejpoužívanějších statistických metod – regresní analýza. Výraz regrese vysvětluje slovník cizích slov [28] jako návrat k dřívějšímu stavu (opak progrese), zpětný postup nebo zpětný vývoj. Tento výraz ve statistice poprvé použil Francis Galton, ke konci 19. století [36], při vyšetřování závislosti výšky synů na výšce otců.

V matematice se závislost hodnot jedné proměnné na hodnotách druhé proměnné vysvětluje funkčním vztahem následovně [17]:

$$y = f(x) \tag{13}$$

Při znalosti konkrétní hodnoty x lze jasně určit, jakou hodnotu bude mít proměnná y. V praktickém životě však na sledovanou veličinu nepůsobí obvykle pouze jedna náhodná veličina X. V mnoha případech je těžké tyto veličiny odhalit a určit jejích přesný vztah ke sledované veličině Y. Pokud jsou ale tyto veličiny závislé, jedná se o závislost stochastickou, nikoli funkční. Veličiny X, Y jsou stochasticky závislé tehdy, pokud změna hodnoty jedné náhodné veličiny vyvolá změnu rozdělení pravděpodobností druhé náhodné veličiny [17].

Regresní funkce umožní předvídat, jakou hodnotu bude mít jedna náhodná veličina, pokud je známá hodnota druhé náhodné veličiny. Cílem regresní analýzy je odhalení tvaru stochastické závislosti a parametrů regresní funkce.

Existují různé typy regresních modelů podle tvaru regresní funkce. Lze rozlišit základní typy [17]:

a) modely lineární vzhledem k parametrům (např. regresní přímka, regresní parabola, regresní hyperbola apod.),

b) modely nelineární vzhledem k parametrům, které je však možné transformovat na lineární (např. regresní mocninná funkce, regresní exponenciální funkce),

c) modely nelineární, které se nedají jednoduše transformovat na lineární tvar.

U prvních dvou typů se odhady parametrů provádějí nejčastěji metodou nejmenších čtverců. U posledního typu tato metoda není vhodná a odhady se provádějí jinými metodami (např. metodou částečných součtů, metodou dílčích průměrů atd.).

4.1.1 Jednoduchý model lineární regrese

Nejjednodušším modelem lineární regrese je takový lineární model, kdy grafem regresní funkce je přímka (tzn. jednoduchý model lineární regrese) [17]. Předpokladem je, že $Y_1, Y_2, ..., Y_n$ je n-tice nekorelovaných náhodných veličin, pro které platí $EY_i = \alpha + \beta x_i$, $DY_i = \sigma^2$, $i = 1, 2, \ldots, n$, kde α, β, σ^2 jsou neznámé parametry, $x_1, x_2, ..., x_n$ je n-tice známých hodnot.

Jednoduchým modelem lineární regrese je tedy model:

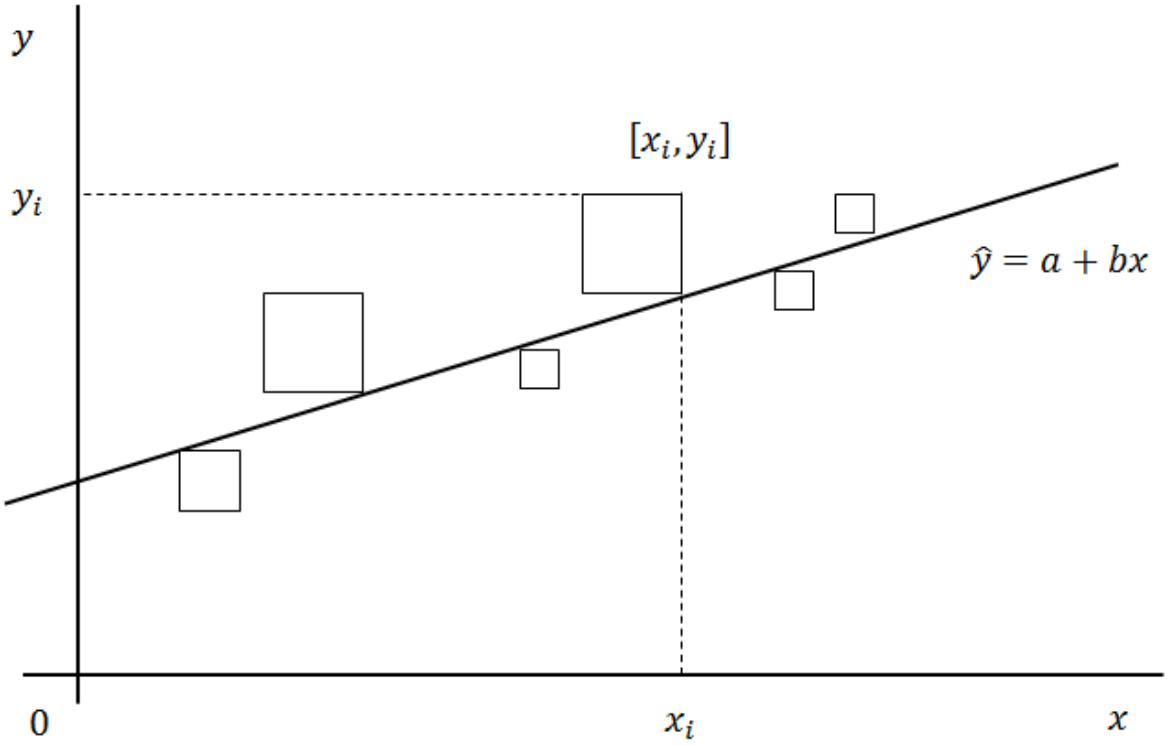
$$Y_i = \alpha + \beta_{x_i} + E_i \tag{14}$$

kde E_i jsou nezávislé náhodné veličiny, pro které platí: $EE_i = 0$, $DE_i = \sigma^2$ a $i = 1, 2, \ldots, n$. E_i je náhodná složka lineárního modelu, která zahrnuje působení náhodných vlivů či působení veličin, které nejsou v modelu zahrnuty.

Jak již bylo zmíněno, nejčastější metodou pro odhad parametrů takového modelu je metoda nejmenších čtverců. Při existenci konkrétních dvojic naměřených hodnot (x_1, y_1), (x_2, y_2), …, (x_n, y_n) se hledá taková funkce $\hat{y} = a + bx$, která by se co nejvíce přibližovala k naměřeným hodnotám. Přiléhání se měří součtem rozdílů hodnoty $\hat{y}_i - y_i$, tzv. reziduí. Odchylky mezi $\hat{y}_i$ a y_i mohou dosahovat kladných, ale i záporných hodnot a při prostém součtu by se tak mohly rozdíly hodnot navzájem odečíst. Míra přiléhání hodnot se tak měří součtem čtverců:

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{15}$$

Z tohoto vzorce (15) je zřejmé, že čím menší tento součet bude, tím bude funkce $\hat{y}$ lépe přiléhat k naměřeným bodům. Cílem je tedy výše zmíněnou funkci minimalizovat. Graficky tuto metodu vysvětluje Obrázek 14.



Obrázek 14: Metoda nejmenších čtverců

Zdroj:vlastní zpracování dle [17]

Metoda nejmenších čtverců poskytuje odhady s optimálními vlastnostmi i při menším počtu pozorování a postup při určení hodnot parametrů je jednoduchý [11]. Tato metoda umožňuje stanovit výběrovou regresní funkci, která vykazuje co nejlepší shodu s napozorovanými daty, kdy kritériem je minimum součtu čtverců reziduí.

Pro zjištění míry shody odhadnutého lineárního modelu s naměřenými daty se nejčastěji používá koeficient vícenásobné determinace [11]. Ten je založen na rozkladu celkového rozptylu vysvětlované proměnné a ukazuje, jakou mírou je vysvětlen rozptyl proměnné *Y* odhadnutým lineárním regresním modelem. Celkový součet čtverců může tak být rozložen na součet čtverců vysvětlený nezávislými proměnnými a na nevysvětlený (reziduální) součet čtverců. Koeficient vícenásobné determinace, označovaný R^2, je podílem vysvětleného součtu čtverců a celkového součtu čtverců (viz vzorec 16).

$$R^2 = \frac{VS\check{C}}{CS\check{C}} = 1 - \frac{NS\check{C}}{CS\check{C}} \tag{16}$$

kde VSČ je vysvětlený součet čtverců, NSČ je nevysvětlený součet čtverců a CSČ je celkový součet čtverců. Koeficient vícenásobné determinace může nabývat hodnot od 0 do 1. V situaci, kdy jsou všechna rezidua nulová, je NSČ nulový a R^2=1. V opačném případě, pokud jsou regresní koeficienty nulové, se NSČ=CSČ a R^2=0.

Jednoduchý model lineární regrese se hodí např. pro již zmíněný vztah výšky otců a synů – tedy pro jednu nezávislou veličinu *X*. V této práci je však zkoumán vliv sentimentu a ostatních faktorů na změnu ceny akcií, proto je třeba využít vícerozměrný model lineární regrese.

4.1.2 Vícerozměrný model lineární regrese

Vícerozměrný model lineární regrese je následující:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + E_i \tag{17}$$

kde $i = 1,2, \ldots, n$, β_i jsou neznámé parametry a E_i jsou nezávislé náhodné veličiny, mající stejné rozdělení pravděpodobnosti se střední hodnotou 0 a kovarianční maticí $\sigma^2\mathbf{I}$, přičemž **I** je jednotková matice [17]. Parametry $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ se odhadují metodou nejmenších čtverců. Funkce $\hat{Y} = B_0 + B_1x_1 + B_2x_2 + \ldots + B_kx_k$ se nazývá výběrová regresní funkce.

4.1.3 Aplikace vícerozměrného modelu lineární regrese

Díky výše zmíněnému vícerozměrnému modelu je možné zkoumat závislost proměnné *Y* (změna ceny akcií) oproti nezávislým *X*. Změna ceny akcií jednotlivých podniků byla vypočtena dle následujícího jednoduchého vzorce:

$$změna\ ceny\ v\ \% = \frac{cena\ po\ zveřejnění - cena\ před\ zveřejněním}{cena\ před\ zveřejněním} * 100 \tag{18}$$

Výchozími údaji pro tento výpočet byly dvě různé ceny akcií. Každá z vybraných bank zveřejnila svou závěrečnou zprávu 10-K k jinému datu a proto byly ceny akcií zjišťovány pro každou z bank k jinému časovému okamžiku. Aby byla získaná data porovnatelná, cena před zveřejněním odkazuje na cenu akcií bank pět dní před zveřejněním jejich závěrečné zprávy. Naopak cena po zveřejnění vyjadřuje cenu akcií pět dnů po zveřejnění zprávy 10-K. Ceny akcií byly vyhledávány především na internetových stránkách NYSE - New Yorské akciové burzy [23].

Mezi nezávislé veličiny *X* patří již zmíněné hodnoty sentimentu a ostatních zbarvených slov (modální slova, neurčitá, právnická) a další ukazatele popsané v podkapitole 3.2. Základní statistiky všech ukazatelů zobrazuje Tabulka 3.

Tabulka 3: Základní statistiky ukazatelů

Proměnná	Popisné statistiky (DATA VŠE UPRAVENO)					
	N platných	Průměr	Medián	Minimum	Maximum	Sm.odch.
LoughranMcDonald Positive	132	0,27	0,24	0,06	0,89	0,16
LoughranMcDonald Uncertainty	132	0,25	0,22	0,06	0,69	0,11
LoughranMcDonald Negative	132	0,25	0,20	0,07	1,11	0,16
LoughranMcDonald Modal Weak	132	0,29	0,24	0,01	0,87	0,21
LoughranMcDonald Modal Strong	132	0,27	0,23	0,00	1,14	0,22
LoughranMcDonald Litigious	132	0,28	0,20	0,03	1,30	0,23
Henry Positive	132	0,30	0,28	0,05	0,83	0,16
Henry Negative	132	0,32	0,27	0,06	1,00	0,18
Tržní kapitalizace (v milionech)	132	2565,96	897,85	53,18	49970,00	5760,05
Objem akcií	132	618231,50	105879,00	429,00	10907000,00	1608374,39
Typ burzy	132	0,81	1,00	0,00	1,00	0,39
P/E	132	17,23	16,03	-5,32	85,55	8,91
P/B	132	1,40	1,29	0,49	3,34	0,44
ROE	132	9,22	8,94	-15,68	86,65	7,76
poměr dluhu a aktiv podniku	132	10,37	8,82	1,17	59,30	6,64

Zdroj: vlastní zpracování v programu STATISTICA

Dalším krokem bylo přezkoumání korelace nezávislých veličin. V Tabulce 4 jsou červeně označeny korelace významné na hladině významnosti α=0,05. Hodnoty korelačního koeficientu dosahují v některých případech i hodnoty kolem 0,8 (např. u negativních a právnických slov podle Loughrana a McDonalda dosahuje koeficient hodnoty 0,83), což by mohlo signalizovat na multikolinearitu.

Multikolinearita je dle [36] případ, kdy sloupce matice nezávislých veličin **X** jsou téměř lineárně závislé. Kromě zmíněných vysokých hodnot výběrových korelačních koeficientů mohou být projevem multikolinearity i nízké hodnoty t statistik u jednotlivých odhadů b_j pro test hypotézy β_j=0, i když koeficient determinace je výrazně nenulový.

Tabulka 4: Korelace mezi nezávislými proměnnými

Proměnná	LoughranMcDonald Positive	LoughranMcDonald Uncertainty	LoughranMcDonald Negative	LoughranMcDonald Modal Weak	LoughranMcDonald Modal Strong	LoughranMcDonald Litigious	Henry Positive	Henry Negative	Tržní kapitalizace (v milionech)	Objem akcií	Typ burzy	P/E	P/B	ROE	poměr dluhu a aktiv podniku
Korelace (DATA VSE UPRAVENO) Označ. korelace jsou významné na hlad. p < ,05000 N=132 (Celé případy vynechány u ChD)															
LoughranMcDonald Positive	1,000	0,608	0,773	0,384	0,304	0,639	0,814	0,655	0,520	0,522	-0,290	-0,016	-0,145	-0,024	0,088
LoughranMcDonald Uncertainty	0,608	1,000	0,700	0,624	0,385	0,650	0,562	0,687	0,374	0,360	-0,335	-0,043	-0,073	-0,025	0,100
LoughranMcDonald Negative	0,773	0,700	1,000	0,379	0,341	0,830	0,672	0,762	0,574	0,622	-0,403	-0,014	-0,056	0,031	0,052
LoughranMcDonald Modal Weak	0,384	0,624	0,379	1,000	0,206	0,275	0,361	0,358	0,134	0,164	-0,151	-0,071	-0,047	-0,066	0,115
LoughranMcDonald Modal Strong	0,304	0,385	0,341	0,206	1,000	0,292	0,295	0,340	0,151	0,212	-0,159	0,131	0,010	-0,001	-0,004
LoughranMcDonald Litigious	0,639	0,650	0,830	0,275	0,292	1,000	0,518	0,595	0,554	0,396	-0,416	-0,017	0,003	0,137	-0,031
Henry Positive	0,814	0,562	0,672	0,361	0,295	0,518	1,000	0,616	0,466	0,402	-0,338	0,042	-0,051	0,014	0,102
Henry Negative	0,655	0,687	0,762	0,358	0,340	0,595	0,616	1,000	0,432	0,507	-0,279	0,016	-0,083	-0,003	0,161
Tržní kapitalizace (v milionech)	0,520	0,374	0,574	0,134	0,151	0,554	0,466	0,432	1,000	0,472	-0,416	-0,064	-0,049	-0,017	0,025
Objem akcií	0,522	0,360	0,622	0,164	0,212	0,396	0,402	0,507	0,472	1,000	-0,270	-0,092	-0,158	-0,079	-0,011
Typ burzy	-0,290	-0,335	-0,403	-0,151	-0,159	-0,416	-0,338	-0,279	-0,416	-0,270	1,000	-0,093	0,059	0,033	-0,075
P/E	-0,016	-0,043	-0,014	-0,071	0,131	-0,017	0,042	0,016	-0,064	-0,092	-0,093	1,000	0,071	-0,182	-0,146
P/B	-0,145	-0,073	-0,056	-0,047	0,010	0,003	-0,051	-0,083	-0,049	-0,158	0,059	0,071	1,000	0,531	-0,247
ROE	-0,024	-0,025	0,031	-0,066	-0,001	0,137	0,014	-0,003	-0,017	-0,079	0,033	-0,182	0,531	1,000	-0,115
poměr dluhu a aktiv podniku	0,088	0,100	0,052	0,115	-0,004	-0,031	0,102	0,161	0,025	-0,011	-0,075	-0,146	-0,247	-0,115	1,000

Zdroj: vlastní zpracování v programu STATISTICA

4.2 Faktorová analýza

Vzhledem k vysoké korelaci nezávislých proměnných tak bylo dalším krokem využití faktorové analýzy. Tato metoda byla vytvořena před více než 70 lety, především k řešení problémů se značným a nepřehledným počtem výchozích proměnných [9]. Cílem metody je nalezení v pozadí stojících, skrytých veličin (faktorů), které vysvětlují závislost proměnných.

Faktorová analýza původně vznikla v psychologii, kdy počátkem 20. století anglický psycholog Charles Spearman ve svém článku o povaze inteligence navrhl hypotézu o existenci společného faktoru obecného intelektu [9]. O rozvoj faktorové analýzy se dále zasloužil L. L. Thurstone, který rozšířil Spearmanův model na vícefaktorový. Po dlouhou dobu byla faktorová analýza využívána především v psychologii. Během posledních desetiletí se však rozšířila i do jiných oborů a v současné době faktorová analýza nachází své uplatnění

v mnoha oblastech [17], např. v ekonomii a ekonometrii (analýza poptávky a nabídky), sociologii (klasifikace postojů, demografických statistik), organizaci a řízení (personální práce – výběr osob pro povolání), medicíně (klasifikace nemocí) atd.

Jedním z cílů faktorové analýzy je posoudit strukturu vztahů sledovaných proměnných a vytvořit tak nové nekorelované proměnné (faktory), které mohou pomoci lépe pochopit analyzovaná data [9]. Dalším cílem je redukce dat, která ale ustupuje před potřebou vysvětlit napozorované korelace pomocí hypotetických faktorů. Faktorová analýza se snaží o co nejlepší reprodukci vzájemných lineárních vztahů původních proměnných a je pokládána za úspěšnou, jestliže se velký počet důsledků podaří objasnit malým počtem příčin [1].

4.2.1 Model faktorové analýzy

Definovat model faktorové analýzy lze podle [9] následovně: nechť $\boldsymbol{x}$ je p-rozměrný náhodný vektor uvažovaných veličin s vektorem středních hodnot $\boldsymbol{\mu}$, s kovariančí maticí $C(\boldsymbol{x}) = \boldsymbol{\Sigma}$ a korelační maticí korelačních koeficientů $P(\boldsymbol{x}) = \boldsymbol{P}$. Model faktorové analýzy předpokládá existenci R v pozadí stojících společných faktorů $F_1, F_2, ...,F_R$, kterých by mělo být výrazně méně než p. Poté lze j-tou pozorovatelnou náhodnou veličinu X_j (j = 1, 2, …, p) vyjádřit následovně:

$$X_j = \mu_j + \gamma_{j1}F_1 + \gamma_{j2}F_2 + \cdots + \gamma_{jR}F_R + \varepsilon_j \tag{19}$$

kde $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ tvoří náhodné (chybové) složky – specifické faktory.

Maticový zápis je následující:

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Gamma f} + \boldsymbol{\varepsilon} \tag{20}$$

kde $\boldsymbol{\Gamma}$ je matice faktorových zátěží typu p x R, $\boldsymbol{f}$ je R-členný vektor společných faktorů a $\boldsymbol{\varepsilon}$ je p-členný vektor specifických faktorů. Pro jakýkoli z uvedených tvarů faktorového modelu se předpokládá, že [9]:

- společné faktory F_r pro r = 1, 2, …, R jsou nezávislé a stejně rozdělené náhodné veličiny, jejichž střední hodnoty jsou nulové a rozptyly jednotkové,
- specifické faktory ε_j, pro j = 1, 2, …, p jsou nezávisle rozdělené náhodné veličiny, jejichž střední hodnota je nulová a pro jejich rozptyl platí $D(\varepsilon_j)=\psi_j$,
- faktory F_r a ε_j jsou pro každou kombinaci r = 1, 2, …, R a j = 1, 2, …, p nezávisle rozdělené náhodné veličiny.

Cílem faktorové analýzy je výpočet faktorové matice. Pro její odhad se využívá řada metod, např. modifikovaná metoda hlavních komponent, metoda maximální věrohodnosti,

Jöreskogova metoda a centroidová metoda [17]. Modifikovaná metoda hlavních komponent vychází z předpokladu, že rozptyl pozorovaných hodnot lze rozložit jednak na rozptyl skutečných hodnot a rozptyl chyb, ale také na rozptyl společných částí (tzv. komunalita), rozptyl specifických částí výsledků (specifita) a rozptyl chyb. Komunalita a specifita tvoří rozptyl skutečných hodnot, specifita a rozptyl chyb tvoří složku zvanou unicita, která je částí rozptylu, která je jedinečná buď díky svým specifickým vlastnostem, nebo důsledkem chyb měření.

Při vytváření modelu faktorové analýzy mohou nastat i různé problémy [1], např. otázka existence modelu: pro všechny systémy znaků X nelze sestrojit model faktorové analýzy, tedy najít takové společné faktory nebo dokázat existenci takových faktorů, které by objasnili existující korelaci mezi různými dvojicemi x_i a x_j, Faktorová analýza též nehodnotí výchozí proměnné podle jejich vzájemné důležitosti, pouze zjednodušeně identifikuje v pozadí stojící veličiny.

4.2.2 Řešení faktorové analýzy

Jednoznačné řešení faktorové analýzy nemusí existovat. Pomocí zavedení transformační matice vzniknou nové faktory, které rovněž vyhovují stanovenému modelu s příslušnými podmínkami řešení [9]. V souvislosti s tím se využívá pojem rotace faktorů, označující výpočetní operace, díky kterým se z dané matice faktorových zátěží získá nová matice. Výraz rotace pochází z geometrického zobrazení transformace faktorových zátěží. Ze stále stejné matice tak lze získat nekonečně mnoho faktorových řešení a problémem se může stát výběr optimálního řešení. Literatura věnující se faktorové analýze i různé statistické programy uvádějí a využívají řadu rotačních algoritmů. Základním rozhodnutím je volba mezi ortogonální (pravoúhlou) rotací, vedoucí k řešení s nekorelovanými faktory a šikmou (kosoúhlou) rotací.

Využitím ortogonální rotace lze matici faktorových zátěží interpretovat jako regresní koeficienty závislosti proměnných na faktorech, ale také zároveň jako korelační koeficienty vztahu mezi proměnnými a faktory. Při šikmé rotaci lze získat matici regresních či korelačních koeficientů a také matici, která ukazuje korelaci mezi výslednými faktory [9]. Snahou faktorové rotace je nalezení souřadnicové soustavy prostoru společných faktorů, která by co nejjednodušeji popisovala proměnné. Každá proměnná by tak měla mít vysoké faktorové zátěže u co nejmenšího počtu společných faktorů a u zbývajících faktorů pouze nízké či středně vysoké zátěže.

Jednou z metod faktorových rotací jsou grafické postupy zavedené L. L. Thurstonem a B. B. Cattellem, které spočívají v optickém prokládání souřadnicových os faktorů skupinami bodů. Nevýhodou těchto metod je však časová náročnost a subjektivita. Snaha odstranit tento problém vedla k návrhům matematizace. Dále jsou uvedena některé kritéria (metody) rotace, včetně nejpoužívanější normalizované varimax rotace. Detailnější informace včetně konkrétních vzorců pro výpočet uvádí další literatura, např. [9].

Varimax

Normalizovaná varimax rotace, navržená H. F. Kaiserem, vybírá takovou transformační matici, aby v jednotlivých sloupcích byl co největší součet rozptylů druhých mocnin faktorových zátěží. Metoda varimax je nejpoužívanější metoda pro rotaci faktorů, produkuje ortogonální faktory a může být využita jako výchozí bod pro šikmé metody rotace.

Quartimax

Principem metody quartimax je funkce, která je součtem čtvrtých mocnin faktorových zátěží. Metoda quartimax produkuje obecný faktor a při využití této metody bývají zátěže zbývajících faktorů nižší než při použití metody varimax.

Orthomax

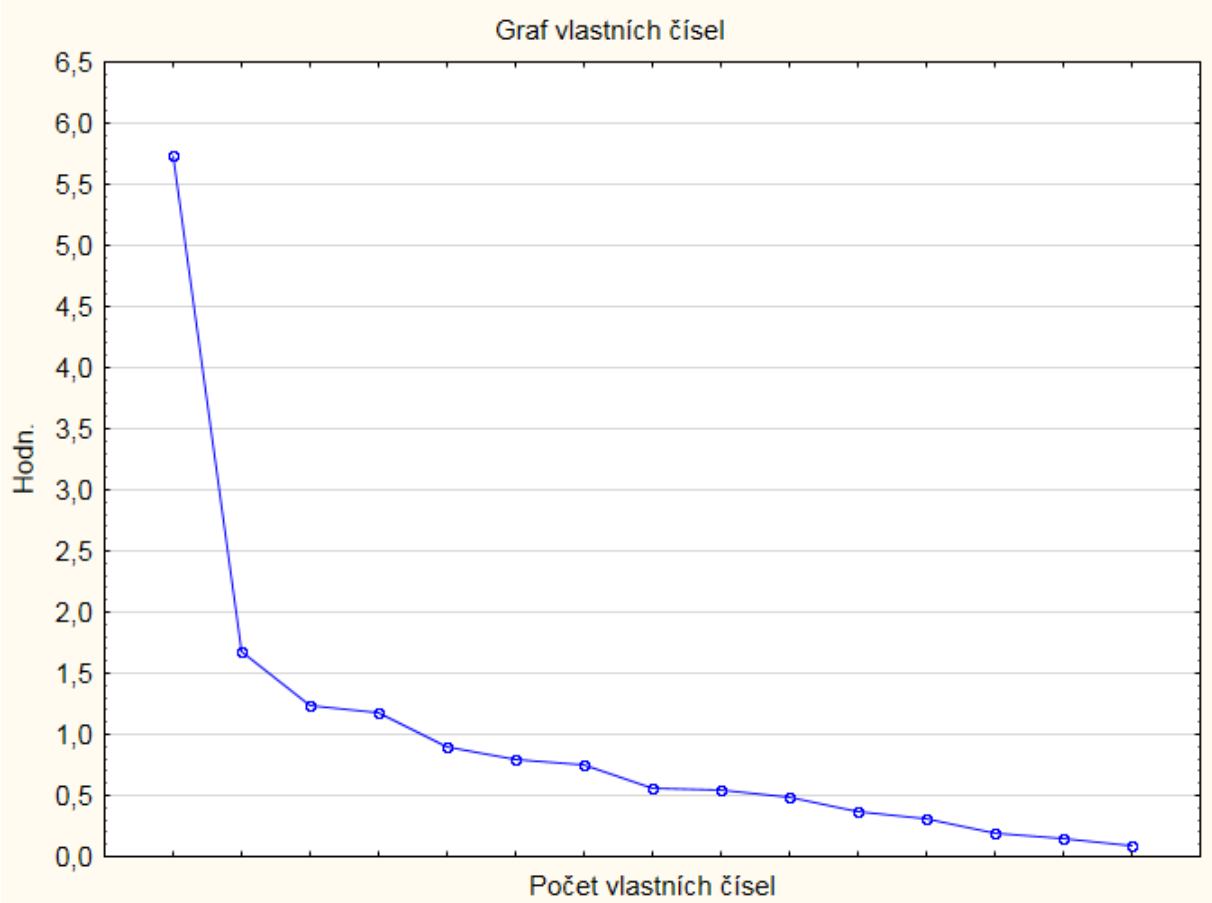
Tato metoda obsahuje skupinu kritérií, která je konstruována jako vážený průměr kritérií quartimax a varimax v obecném maximalizovaném tvaru. Kritérium orthomax obsahuje veličinu δ, která může nabývat hodnot od 0 do 1. V situaci, kdy $\delta = 0$ je kritérium shodné s quartimax kritériem a pro $\delta = 1$ naopak s varimax kritériem. Když $\delta = 0{,}5$ je kritérium označováno biquartimax a při situaci $\delta = R/2$ se jedná o metodu equamax.

Ortogonální transformace bývají preferovány především díky nekorelovaným obecným faktorům. Naopak zastánci šikmých transformací uvádějí, že nezávislé faktory v jedné situaci mohou být závislé v jiné [9]. V některých případech též není možné, aby rotované osy procházely každým shlukem proměnných a zůstaly u toho pravoúhlé. Šikmé rotace se hodí právě pro takovéto situace. Statistické programy nabízejí velký počet postupů, které vedou k šikmým faktorům – např. oblitin, oblimax, biquartimin, promax apod.

4.2.3 Využití faktorové analýzy

Faktorová analýza byla provedena opět pomocí programu STATISTICA. Jako proměnné bylo vybráno všech 15 nezávislých veličin. Dalším krokem byla volba počtu faktorů. Odhad tohoto počtu byl proveden na základě sutinového grafu (viz Obrázek 15), který má charakteristický tvar klesající křivky [4]. Vhodný počet faktorů lze odhadnout podle

největšího poklesu mezi dvěma po sobě následujícími faktory. V tomto případě je nejnižší pokles mezi 1. a 2. faktorem – takový počet faktorů by však byl příliš nízký a nedostačující. Další výrazný pokles nastává mezi 4. a 5. faktorem, proto je tedy v tomto případě počet faktorů stanoven na 4.



Obrázek 15: Sutinový graf

Zdroj: vlastní zpracování v programu STATISTICA

Další možností pro odhad faktorů by mohlo být využití metody hlavních komponent – počet faktorů stanovených touto metodou ale nemusí být konečný. Počet faktorů může být upraven na základě různých subjektivních i objektivních rad a kritérií. Jednou z nich je skutečnost, že do konečného řešení by neměly být zahrnuty triviální faktory – tedy faktory, které korelují jen s jednou z *p* proměnných [9].

Po stanovení počtu faktorů přišel na řadu výběr optimálního řešení, které bylo vyhledáno pomocí různých rotací faktorů. Program STATISTICA umožňuje všechny typy rotací zmíněné v předchozí části textu (4.2.2) – varimax, quartimax, biquartimax a equamax (a to jak prosté, tak normalizované). Ze všech osmi otestovaných variant byla vybrána ta nejčastěji používaná - rotace varimax normalizovaný. Tato rotace byla vybrána především díky

jednoznačné charakteristice jednotlivých faktorů. Výsledky ostatních druhů rotací obsahuje příloha A.

Rotaci varimax normalizovaný zobrazuje následující Tabulka 5, z které lze vyčíst korelace nově vytvořených faktorů s jednotlivými proměnnými. První faktor je vysvětlován především sentimentálními slovy – pozitivními, negativními a právnickými (*LoughranMcDonald Positive*, *LoughranMcDonald Negative*, *LoughranMcDonald Litigious*). Vliv na něj mají ale i další proměnné - tržní kapitalizace a objem akcií. Všechny tyto nezávislé proměnné mají na faktor přímý vliv.

Faktor 1 je ze všech faktorů vysvětlován celkem nejvíce proměnnými, především sentimentálními slovy, které mezi sebou významně korelovaly (viz Tabulka 4). Kromě červeně znázorněných výše zmíněných významných vztahů mají vliv na první faktor i pozitivní a negativní slova podle Elaine Henry (*Henry Positive*, *Henry Negative*). V další části této práce bude faktor 1 označován jako **sentiment a velikost**.

Tabulka 5: Faktorové zátěže

Proměnná	Faktor. zátěže (Varimax normaliz.) (DATA VŠE UPRAVENO) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)			
	Faktor (1)	Faktor (2)	Faktor (3)	Faktor (4)
LoughranMcDonald Positive	0,710754	0,059979	0,041193	0,479745
LoughranMcDonald Uncertainty	0,457968	0,005211	0,041651	0,759466
LoughranMcDonald Negative	0,808976	-0,045520	0,006992	0,459018
LoughranMcDonald Modal Weak	0,047958	0,056043	0,138825	0,797186
LoughranMcDonald Modal Strong	0,163256	-0,031386	-0,355549	0,530567
LoughranMcDonald Litigious	0,741806	-0,180314	-0,027810	0,357765
Henry Positive	0,623110	0,008596	-0,019125	0,491599
Henry Negative	0,613323	0,034140	0,033012	0,553471
Tržní kapitalizace (v milionech)	0,810415	0,001748	0,047630	-0,018300
Objem akcií	0,725415	0,125631	0,069267	0,061748
Typ burzy	-0,570794	-0,074184	0,159192	-0,040719
P/E	-0,054207	0,129143	-0,862070	0,056297
P/B	-0,117194	-0,837420	-0,164122	0,012733
ROE	0,012647	-0,856365	0,209323	-0,001009
poměr dluhu a aktiv podniku	-0,056645	0,389207	0,502392	0,243538
Výkl.roz	4,241855	1,667657	1,250120	2,679094
Prp.celk	0,282790	0,111177	0,083341	0,178606

Zdroj: vlastní zpracování v programu STATISTICA

Druhý faktor je vysvětlován ukazateli P/B a ROE (poměrem tržní a účetní ceny akcie a rentabilitou vlastního kapitálu). Korelační koeficient mezi těmito proměnnými dosahoval hodnoty zhruba 0,53. Obě proměnné s druhým faktorem významně korelují, tentokrát se však jedná o nepřímý vztah. Kromě těchto P/B a ROE nemá na druhý faktor už žádná jiná proměnná vliv. Druhý faktor bude dále označován jako **růst podniku**.

Faktor 3 je vysvětlován pouze jednou proměnnou - ukazatelem P/E (poměr tržní ceny akcie a zisku na akcii), kdy se jedná opět o nepřímou závislost. Tento faktor bude označen prostě **P/E**. Poslední, čtvrtý, faktor je vysvětlen především neurčitými a slabě modálními slovy. Hodnota korelačního koeficientu těchto dvou proměnných byla zhruba 0,62. Čtvrtý faktor bude v dalším textu označen jako **neurčitost**.

Před provedením regresní analýzy byla pro kontrolu ověřena korelace nově vytvořených faktorů. Jak ukazuje Tabulka 6, tentokrát jsou hodnoty již nekorelované.

Tabulka 6: Korelace mezi vytvořenými faktory

Proměnná	Korelace (Faktor. skóre (DATA VŠE UPRAVENO) v PS1) Označ. korelace jsou významné na hlad. p < ,05000 N=132 (Celé případy vynechány u ChD)			
	Sentiment a velikost	Růst podniku	P/E	Neurčitost
Sentiment a velikost	1,000000	0,000000	0,000000	0,000000
Růst podniku	0,000000	1,000000	0,000000	0,000000
P/E	0,000000	0,000000	1,000000	0,000000
Neurčitost	0,000000	0,000000	0,000000	1,000000

Zdroj: vlastní zpracování v programu STATISTICA

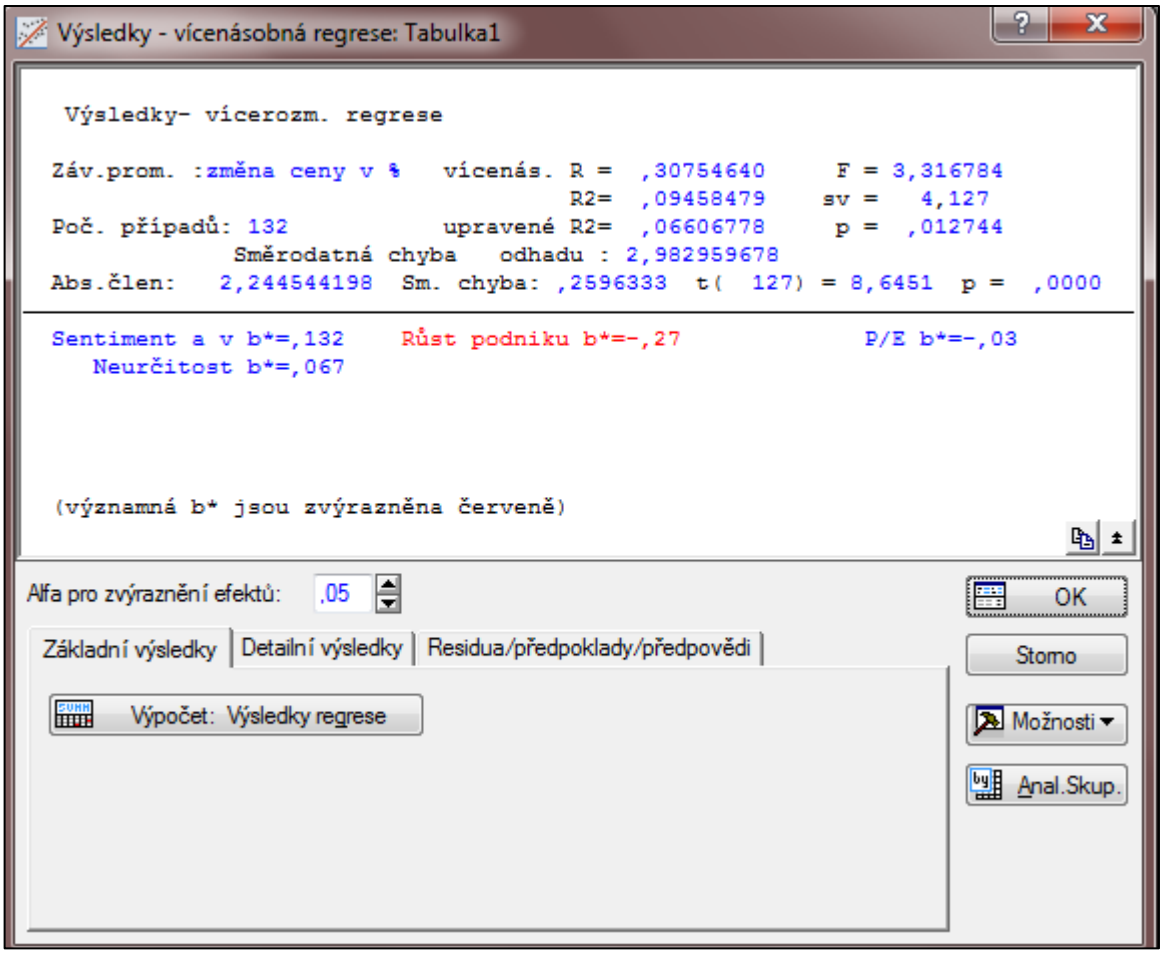
4.3 Regresní analýza s využitím faktorů

S využitím faktorové analýzy byly vytvořeny čtyři umělé faktory, které mezi sebou nejsou korelované (na rozdíl oproti původním proměnným). Závislou veličinou byla opět považována změna ceny akcií podniků, nezávislými veličinami byly nově vytvořené faktory.

Z následujícího Obrázku 16 a Tabulky 7 vyplývá, že na hladině významnosti α=0,05 je významný pouze faktor **Růst podniku** s *p*-hodnotou zhruba 0,002 (na obrázcích označen červeně). Jak již bylo zmíněno v předchozí části textu (4.2.3), tento je vysvětlen především hodnotou ROE (rentability vlastního kapitálu) a ukazatelem P/B (poměr tržní a účetní ceny akcie). Faktor Růst podniku ovlivňuje změnu ceny přímo úměrně, přestože má hodnotu regresního koeficientu *b**= -0,27. To je dáno opačným znaménkem u faktorových zátěží (spíše než o růstu by se tedy dalo hovořit o poklesu podniku).

Druhou nejvyšší hodnotu standardizovaného koeficientu *b* (*b**) má první faktor (**Sentiment a velikost**), který na hladině α=0,05 není statisticky významný. S *p*-hodnotou zhruba 0,12 se blíží hladině významnosti α=0,1. Faktor Sentiment a velikost má pozitivní vliv na růst cen akcií podniků.

Poslední dva faktory, týkající se P/E a neurčitých a slabých modálních slov, mají na změnu ceny akcií podniků pouze velmi slabý vliv. Jejich *p*-hodnoty dosahují vysokých hodnot (0,43 a 0,76). V případě třetího faktoru (**P/E**) je vliv na změnu ceny nepřímo úměrný. U posledního faktoru (**Neurčitost**) je vliv přímo úměrný.



Obrázek 16: Kvalita vícenásobné regrese

Zdroj: vlastní zpracování v programu STATISTICA

Tabulka 7: Koeficienty vícenásobné regrese

N=132	Výsledky regrese se závislou proměnnou : změna ceny v % (Tabulka1) R= ,30754640 R2= ,09458479 Upravené R2= ,06606778 F(4,127)=3,3168 p<,01274 Směrod. chyba odhadu : 2,9830					
	b*	Sm.chyba	b	Sm.chyba	t(127)	p-hodn.
Abs.člen			2,244544	0,259633	8,64505	0,000000
Sentiment a velikost	0,132182	0,084435	0,408002	0,260622	1,56549	0,119956
Růst podniku	-0,268280	0,084435	-0,828092	0,260622	-3,17736	0,001866
P/E	-0,025618	0,084435	-0,079073	0,260622	-0,30340	0,762081
Neurčitost	0,066949	0,084435	0,206649	0,260622	0,79291	0,429311

Zdroj: vlastní zpracování v programu STATISTICA

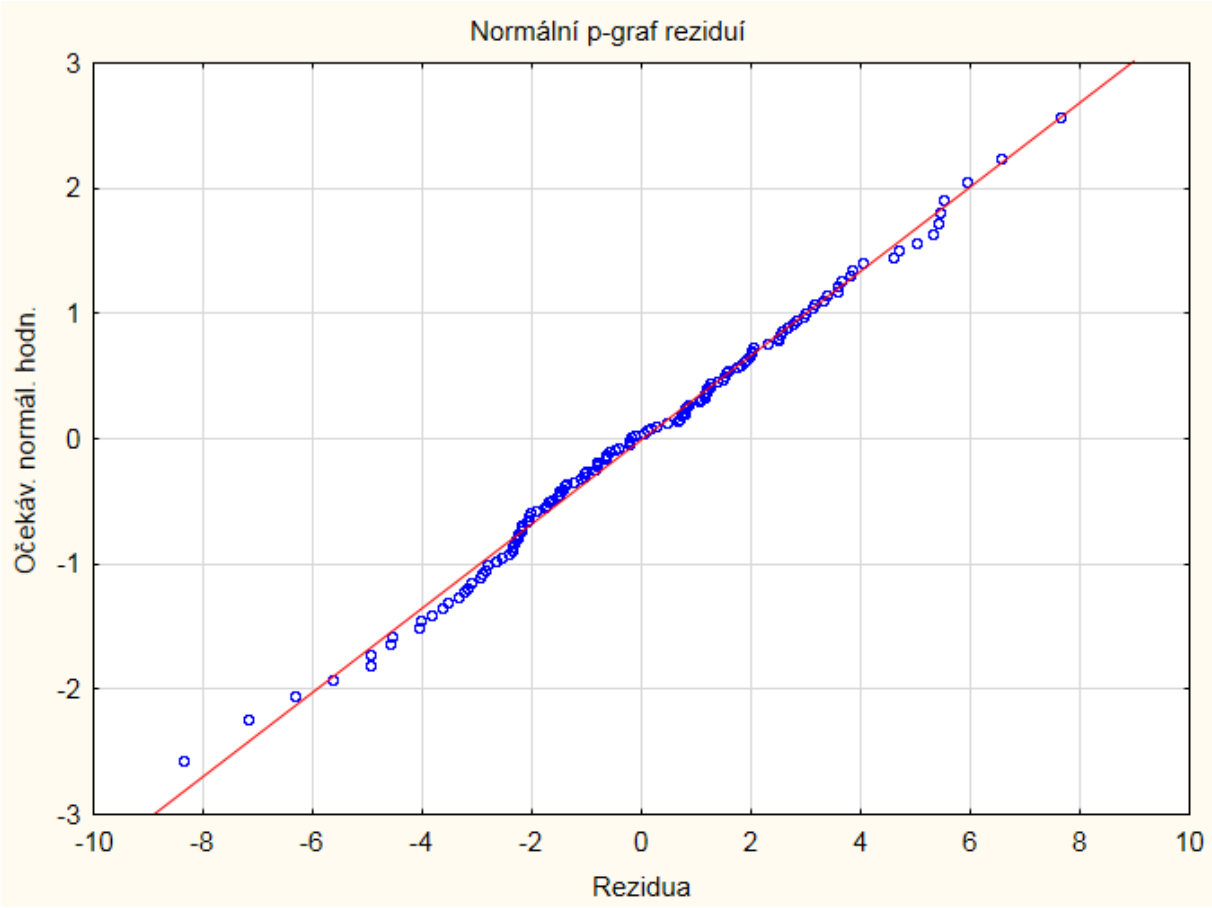
4.3.1 Ověření předpokladů modelu

Cílem této práce bylo zjistit, zda má sentiment vliv na změnu ceny akcií. Hlavním cílem tedy nebylo sestavení modelu pro předpověď změny ceny, ani zajištění vysokého podílu vysvětlené variability. Z nízké hodnoty koeficientu R^2 (9,46 % - viz Obrázek 16) lze usoudit, že na změnu ceny akcií mají silný vliv i jiné faktory, které v této práci nebyly uvažovány.

Kromě hodnocení významnosti parametrů modelu a koeficientu determinace je možné ověřovat i další předpoklady lineárního modelu – mimo linearity závislosti i předpoklady o chybách: normalita, homoskedasticita a nezávislost jednotlivých pozorování [25]. Linearitou se předpokládá, že malá změna nezávislé proměnné změní hodnotu závislé proměnné pouze minimálně. Pokud však změna nezávislé proměnné bude větší, tím větší bude i očekávaná změna závislé hodnoty.

Normalita může být ověřena testováním normality odchylek modelu od skutečnosti (testováním normality reziduí). K ověření předpokladů na základě reziduí slouží v programu STATISTICA záložka *Rezidua/předpoklady/předpovědi* (v probíhající regresní analýze) a následně možnost *Reziduální analýza.*

Jednou z možností, které reziduální analýza nabízí, je *Normální p-graf reziduí*. Tímto grafem si lze ověřit normalitu reziduí. V případě normality by se body reziduí měly nacházet na přímce nebo okolo ní [29]. Jak ukazuje Obrázek 17, předpoklad normality reziduí byl splněn, pro kontrolu však byl proveden i test normality.

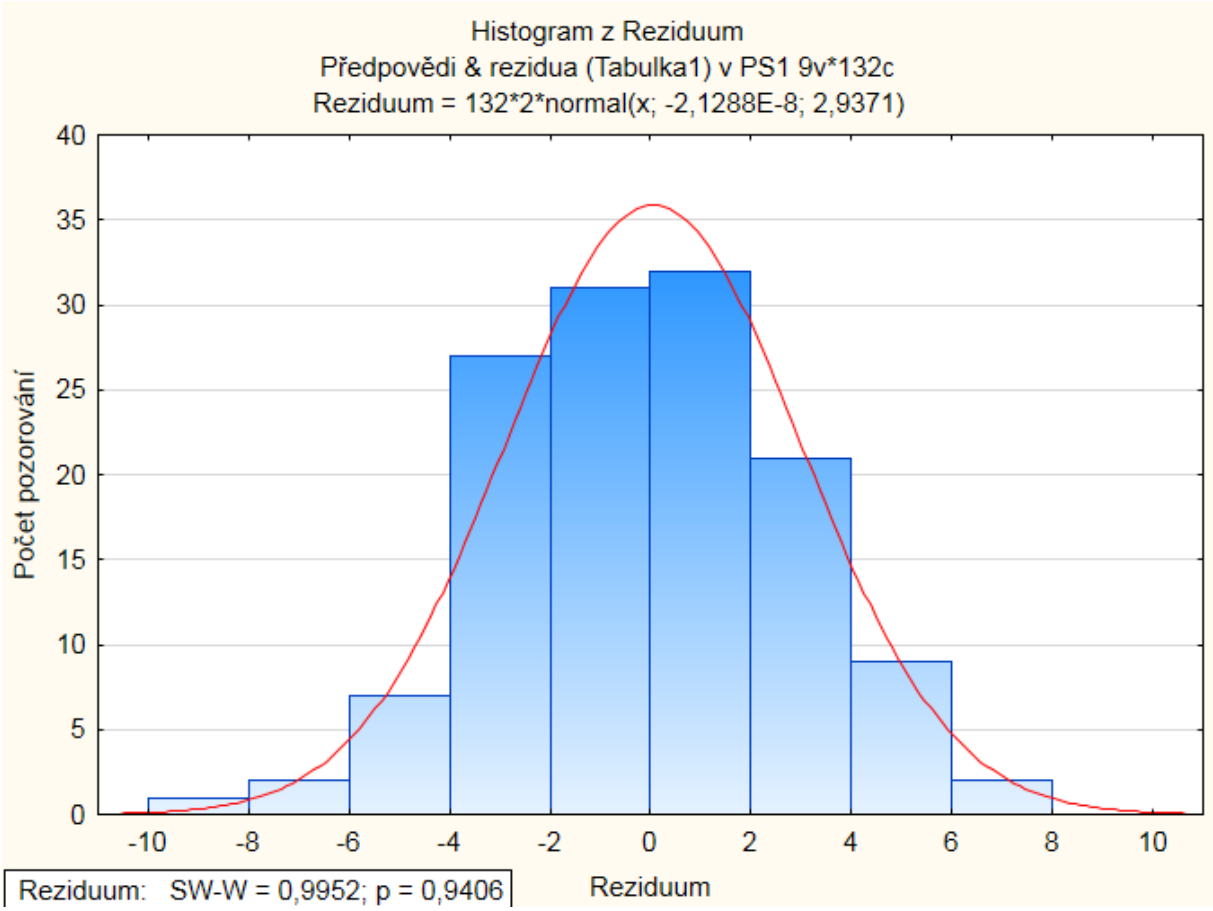


Obrázek 17: Normální p-graf reziduí

Zdroj: vlastní zpracování v programu STATISTICA

Vypočtená rezidua pro jednotlivé podniky byla využita jako aktivní vstup pro histogram, proměnnou byla samozřejmě rezidua a vybrán Shapiro-Wilkův test pro ověření normality. Při pohledu na histogram (Obrázek 18) je zřejmé, že rozdělení není dokonale symetrické, avšak odpovídá normálnímu rozdělení.

Výsledek Shapiro-Wilkova testu je poté naprosto zřejmý: *p*-hodnota (0,9406) je vyšší než hladina významnosti α=0,05. Nulová hypotéza se tedy nezamítá – rezidua mohou mít normální rozdělení pravděpodobnosti.

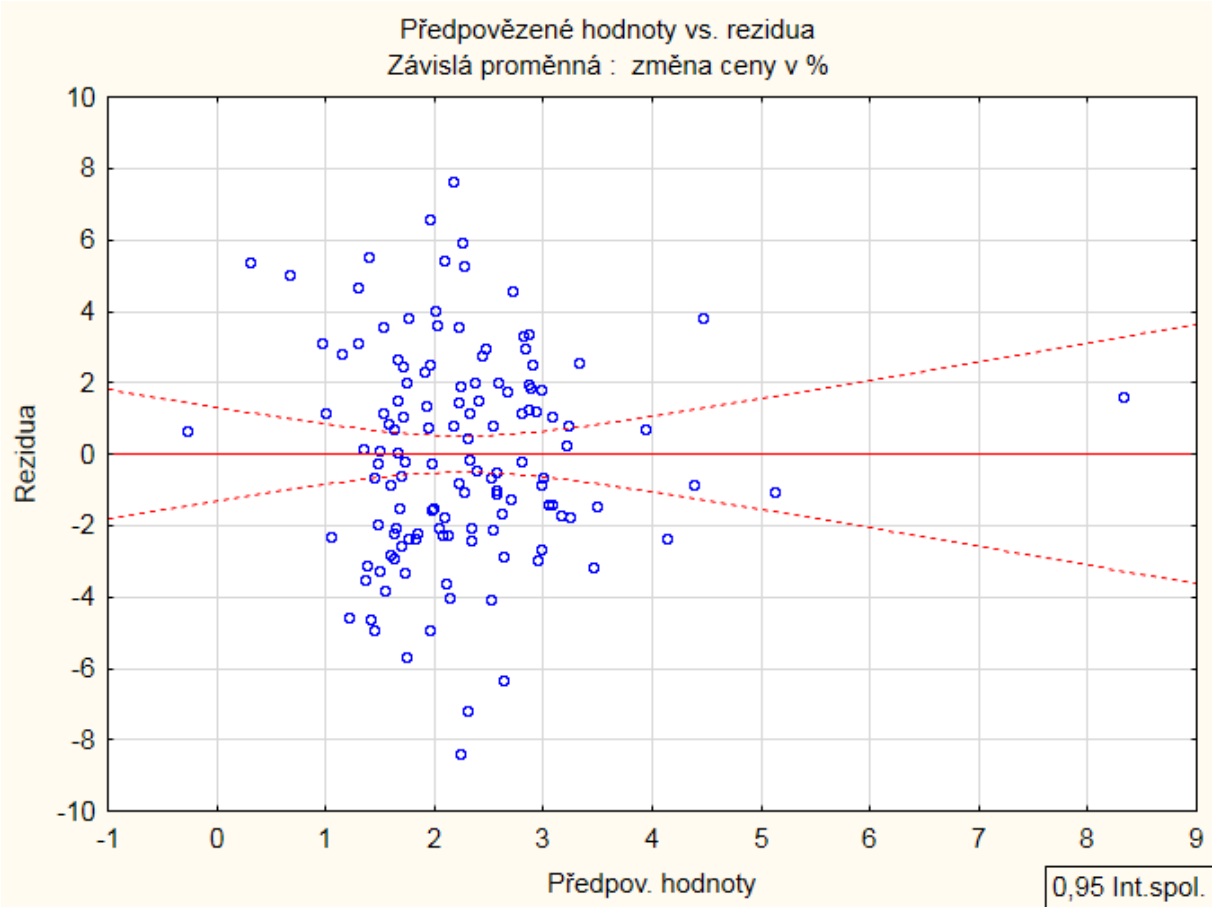


Obrázek 18: Histogram z vypočtených reziduí

Zdroj: vlastní zpracování v programu STATISTICA

Dalším předpokladem regresního modelu je, že střední hodnota chybové složky je rovna 0 a chybová složka má konstantní rozptyl (homoskedasticita) [29]. K ověření tohoto předpokladu lze využít další z možností reziduální analýzy - bodový graf *Předpovědi vs. rezidua*.

Z následujícího Obrázku 19 jasně vyplývá, že chybová složka má konstantní rozptyl (podmínka homoskedasticity). Rezidua modelu jsou zhruba stejně rozptýlena kolem nulové střední hodnoty. Správnost modelu se dá vysvětlit i tak, že po celé délce x-ové osy, jsou rezidua rovnoměrně zastoupena jak nad nulou tak i pod ní [25].



Obrázek 19: Rozptyl reziduí

Zdroj: vlastní zpracování v programu STATISTICA

ZÁVĚR

Cílem této diplomové práce bylo shrnutí současných možností analýzy sentimentu, provedení sběru dat a analýzy sentimentu ve výročních zprávách a analyzování závislostí mezi sentimentem a vývojem podniků na kapitálových trzích. Mezi současné možnosti analýzy sentimentu lze zařadit např. hodnocení recenzí (tedy názorů na určitý produkt). Podniky chtějí zjistit názory svých stávajících nebo potenciálních zákazníků na své produkty a služby. Stejně tak zákazníky zajímá, zda si produkt někdo zakoupil a jak jej hodnotí. Dříve byly tyto názory získávány především pomocí průzkumu trhu. V současné době však s růstem informací zveřejňovaných na internetu lze využít například analýzy sentimentu.

Další možností analýzy sentimentu je její využití v oblasti financí. V této oblasti byly v posledních letech zkoumány dva typy sentimentu – sentiment investorů a sentiment založený na textu. Mezi zdroje sentimentu patří sentiment vyjadřovaný podniky, sentiment vyjadřovaný v médiích a sentiment vyjadřovaný na internetu. Výhodou dokumentů zveřejňovaných podniky jsou informace „z první ruky“, není však jisté, zda management zveřejní pouze pravdu. Další nevýhodou je, že tyto dokumenty bývají zveřejňovány pouze jednou ročně nebo čtvrtletně. Zprávy v médiích a online média zveřejňují své informace častěji, ne vždy se však věnují konkrétním podnikům.

V oblasti financí byl v posledních letech zkoumán vliv sentimentu na různé proměnné či události. Někteří autoři se zaměřili na vliv sentimentu na ceny a výnosy cenných papírů. Např. Loughran a McDonald s využitím svého vlastního slovníku sentimentu prokázali vliv negativních slov na výnosy z cenných papírů. Vliv sentimentu byl prokázán i na objem tržních obchodů. Analýzu sentimentu je možné využít i pro zvýšení přesnosti predikce finančních podvodů a bankrotů.

Pro zkoumání vlivu sentimentu na vývoj podniků na kapitálových trzích bylo v této práci vybráno 132 amerických bank. Vývoj podniků na kapitálových trzích byl reprezentován procentní změnou ceny akcie. Hodnocení sentimentu bylo provedeno na základě dvou různých slovníků sentimentu (prvním byl slovník pozitivních a negativních slov podle Elaine Henry, druhým byl slovník pozitivních, negativních, neurčitých, právnických, slabě a silně modálních slov dle Loughrana a McDonalda). S využitím inverzní četnosti výskytu slov v dokumentu byla zjištěna průměrná hodnota sentimentu jednotlivých podniků. K těmto průměrným hodnotám sentimentu byly vybrány další ukazatele – tržní kapitalizace, objem

akcií obchodovaných na burze, typ burzy, P/E, P/B, ROE a poměr celkového dluhu k celkovým aktivům.

Metodou pro zjištění vlivu sentimentu na změnu ceny akcií byla jedna z nejpoužívanějších statistických metod – regresní analýza, konkrétně vícerozměrná regresní analýza. Závislou proměnnou byla zvolena procentní změna ceny akcie, nezávislé proměnné byly hodnoty sentimentu a ostatní ukazatele. Některé nezávislé veličiny však mezi sebou silně korelovaly, proto bylo třeba využít faktorovou analýzu. Pomocí této metody byly vytvořeny čtyři nové faktory, které mezi sebou již nekorelovaly. Na základě těchto faktorů byla poté provedena vícerozměrná regresní analýza.

Na hladině významnosti α=0,05 byl významný pouze jeden faktor – označený jako Růst podniku, který je vysvětlen především hodnotou ROE a P/B. Třetí faktor (P/E) a čtvrtý faktor (Neurčitost) mají na změnu ceny akcií podniků pouze malý vliv.

Vliv sentimentu byl charakterizován především prvním faktorem – Sentiment a velikost. Na hladině významnosti α=0,05 nebyl tento vliv statisticky významný. Při stanovení hladiny významnosti α=0,1 by byl vliv sentimentu téměř významný, protože jeho *p*-hodnota dosahovala hodnoty přibližně 0,12. Vliv sentimentu na vývoj cen akcií bank na kapitálových trzích tedy není zanedbatelný. Růst relativního počtu zabarvených slov (sentimentu) v textu výročních zpráv bank vede k růstu cen jejich akcií. Na závěr byly ověřeny předpoklady regresního modelu. Cíl práce byl takto splněn.

POUŽITÁ LITERATURA

[1] AJVAZJAN, S. A., BEŽAJEVA, Z. I., STAROVEROV, O. V. *Metody vícerozměrné analýzy*. 1. vyd. Překlad Jiří Hustopecký. Praha: SNTL - Nakladatelství technické literatury, 1981. 252 s.

[2] CECCHINI, M., AYTUG, H., KOEHLER, G. J., PATHAK, P. Making Words Work: Using Financial Text as a Predictor of Financial Events. *Decision Support Systems*. 2010, roč. 50, č. 1, s. 164-175. ISSN 1873-5797.

[3] DATOVÁ ŽURNALISTIKA. *Analýza sentimentu: Barometr nálady* [online]. 2010 [cit. 2015-04-11]. Dostupné na: <http://www.datovazurnalistika.cz/analyza-sentimentu-barometr-nalady/>.

[4] FIELD, Andy P. *Discovering statistics using SPSS for Windows: advanced techniques for the beginner.* 1. vyd. Thousand Oaks: Sage Publications, 2000. 496 s. ISBN 07-619-5755-3.

[5] FOREX FRIENDS. *Analýza sentimentu trhu* [online]. 2012 [cit. 2015-04-11]. Dostupné na: < http://www.forexfriends.cz/3-32-47-pruvodce-.aspx>.

[6] GHOSE, A., IPEIROTIS, P. G. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*. 2011, roč. 23, č. 10, s. 1498-1512. ISSN 1041-4347.

[7] HÁJEK, P., OLEJ, V. Evaluating Sentiment in Annual Reports for Financial Distress Prediction Using Neural Networks and Support Vector Machines. *Communications in Computer and Information Science*. 2013, č. 384, s. 1-10. ISSN 1865-0929.

[8] HÁJEK, P., OLEJ, V., MYŠKOVÁ, R. Forecasting Corporate Financial Performance Using Sentiment in Annual Reports for Stakeholders' Decision-Making. *Technological and Economic Development of Economy*. 2014, roč. 20, č. 4, s. 721-738. ISSN 2029-4921.

[9] HEBÁK, P., HUSTOPECKÝ, J., PECÁKOVÁ, I., PRŮŠA, M., ŘEZANKOVÁ, H., SVOBODOVÁ, A., VLACH, P. *Vícerozměrné statistické metody*. 2. přeprac. vyd. Praha: Informatorium, 2007. 271 s. 3. sv. ISBN 978-80-7333-001-9.

[10] HENRY, E. Are Investors Influenced By How Earnings Press Releases Are Written? *Journal of Business Communication*. 2008, roč. 45, č. 4, s. 363-407. ISSN 2329-4892.

[11] HUŠEK, R. *Ekonometrická analýza*. 1. vyd. Praha: Oeconomica, 2007. 367 s. ISBN 978-80-245-1300-3.

[12] INVESTOPEDIA. *10-K* [online]. [cit. 2015-04-11]. Dostupné na: <http://www.investopedia.com/terms/1/10-k.asp>.

[13] INVESTOPEDIA. *Management Discussion and Analysis – MD&A* [online]. [cit. 2015-04-11]. Dostupné na: <http://www.investopedia.com/terms/m/mdanalysis.asp>.

[14] KEARNEY, C., LIU, S. Textual Sentiment in Finance: A Survey of Methods and Models. *International Review of Financial Analysis.* 2014, roč. 23, č. 33, s. 171-185. ISSN 1057-5219.

[15] KIM, S., PANTEL, P., CHKLOVSKI, T., PENNACCHIOTTI, M. Automatically assessing review helpfulness. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*. Morristown, NJ, USA: Association for Computational Linguistics, 2006, s. 423-430. ISBN 1932432736. DOI: 10.3115/1610075.1610135. Dostupné na: <http://portal.acm.org/citation.cfm?doid=1610075.1610135>.

[16] KNÁPKOVÁ, A., PAVELKOVÁ, D., ŠTEKER, K. *Finanční analýza: komplexní průvodce s příklady*. 2. rozš. vyd. Praha: Grada, 2013. 236 s. ISBN 978-80-247-4456-8.

[17] KUBANOVÁ, J. *Statistické metody pro ekonomickou a technickou praxi*. 3. dopl. vyd. Bratislava: Statis, 2008. 247 s. ISBN 978-80-85659-47-4.

[18] LIU, B. *Sentiment Analysis and Opinion Mining*. 1. vyd. San Rafael: Morgan & Claypool Publishers, 2012. 180 s. Synthesis lectures on human language technologies, 16. ISBN 978-1608458844.

[19] LOUGHRAN, T., MCDONALD, B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*. 2011, roč. 66, č. 1, s. 35-65. ISSN 1540-6261.

[20] LU, Y., TSAPARAS, P., NTOULAS, A., POLANYI, L. Exploiting social context for review quality prediction. In: *Proceedings of the 19th international conference on World wide web - WWW '10*. New York, New York, USA: ACM Press, 2010, s. 691-700. ISBN 9781605587998. DOI: 10.1145/1772690.1772761. Dostupné na: <http://research.microsoft.com/pubs/120365/fp256-lu.pdf.>.

[21] MARKET WATCH. [online]. [cit. 2015-04-11]. Dostupné na: < http://www.marketwatch.com/>.

[22] MINER, G., ELDER, J., FAST, A., HILL, T., NISBET, R., DELEN, D. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. 1. vyd. Waltham: Academic Press, 2012. 1053 s. ISBN 978-0123869791.

[23] NEW YORK STOCK EXCHANGE. [online]. [cit. 2015-04-11]. Dostupné na: < https://www.nyse.com/index>.

[24] POLOUČEK, S. a kol. *Peníze, banky, finanční trhy*. 1. vyd. Praha: C. H. Beck, 2009. 415 s. ISBN 978-80-7400-152-9.

[25] PROCHÁZKA, B. *Stručná biostatistika pro lékaře*. 1. vyd. Praha: Karolinum, 2015. 125 s. ISBN 97-880-246-27830.

[26] REJNUŠ, O. *Finanční trhy*. 4. aktualiz. a rozš. vyd. Praha: Grada, 2014. 760 s. ISBN 978-80-247-3671.

[27] SAXOBANK. *Volatilita* [online]. [cit. 2015-04-11]. Dostupné na: < http://cz.saxobank.com/support/slovnik-pojmu/volatilita>.

[28] SLOVNÍK CIZÍCH SLOV. [online]. [cit. 2015-04-11]. Dostupné na: <http://slovnik-cizich-slov.abz.cz>.

[29] STATSOFT. *Úvod do regresní analýzy* [online]. 2014 [cit. 2015-04-11]. Dostupné na: < http://www.statsoft.cz/file1/PDF/newsletter/2014_26_03_StatSoft_Uvod_do_regresni_analyzy.pdf>.S

[30] TETLOCK, P. C., SAAR-TSECHANSKY, M., MACSKASSY, S. More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*. 2008, roč. 63, č. 3, s. 1437-1467. ISSN 1540-6261.

[31] TETLOCK, P. C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*. 2007, roč. 62, č. 3, s. 1139-1168. ISSN 1540-6261.

[32] U. S. SECURITIES AND EXCHANGE COMMISSION. *EDGAR: Company Filings* [online]. [cit. 2015-04-11]. Dostupné na: < http://www.sec.gov/edgar/searchedgar/companysearch.html>.

[33] UNIVERSITY OF NOTRE DAME. *Bill McDonald* [online]. [cit. 2015-04-11]. Dostupné na: < http://www3.nd.edu/~mcdonald/Word_Lists.html>.

[34] WEISS, S. M., INDURKHYA N., ZHANG, T. *Fundamentals of Predictive Text Mining*. 1. vyd. New York: Springer, 2010. 226 s. ISBN 978-1849962254.

[35] YORICK, W., BIEN, J. Beliefs, Points of View, and Multiple Environments. *Cognitive Science*. 1983, roč. 7, č. 2, s. 95-119. ISSN 1551-6709.

[36] ZVÁRA, K. *Regresní analýza*. 1. vyd. Praha: Academia, 1989. 245 s. ISBN 80-200-0125-5.

SEZNAM PŘÍLOH

Příloha A

Varianta bez rotace

Proměnná	Faktor. zátěže (Bez rot.) (DATA VŠE UPRAVENO) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)			
	Faktor (1)	Faktor (2)	Faktor (3)	Faktor (4)
LoughranMcDonald Positive	-0,859319	-0,035174	0,021075	-0,022604
LoughranMcDonald Uncertainty	-0,810579	-0,008676	0,127727	0,338909
LoughranMcDonald Negative	-0,924059	0,081403	-0,013208	-0,081087
LoughranMcDonald Modal Weak	-0,499069	-0,102736	0,282171	0,566546
LoughranMcDonald Modal Strong	-0,432423	0,076964	-0,246977	0,426197
LoughranMcDonald Litigious	-0,806243	0,218746	-0,034783	-0,111673
Henry Positive	-0,791937	0,018360	-0,011894	0,052277
Henry Negative	-0,820769	-0,017758	0,047888	0,092088
Tržní kapitalizace (v milionech)	-0,653998	0,046316	-0,082484	-0,471929
Objem akcií	-0,634512	-0,087740	-0,052828	-0,370846
Typ burzy (0= NYSE, 1=Nasdaq)	0,491536	0,011617	0,242414	0,240420
P/E Current	0,017673	-0,000526	-0,821656	0,300865
Price to Book Ratio	0,118108	0,841804	0,003963	0,139660
Return on Equity	0,016015	0,812366	0,339730	-0,041554
Total Debt to Total Assets	-0,111494	-0,473413	0,472704	0,079997
Výkl.roz	5,735140	1,675590	1,243553	1,184443
Prp.celk	0,382343	0,111706	0,082904	0,078963

Varianta varimax prostý

Proměnná	Faktor. zátěže (Varimax pr.) (DATA VŠE UPRAVENO) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)			
	Faktor (1)	Faktor (2)	Faktor (3)	Faktor (4)
LoughranMcDonald Positive	0,792435	0,079143	-0,014319	0,325882
LoughranMcDonald Uncertainty	0,605965	0,005474	0,011729	0,648787
LoughranMcDonald Negative	0,888168	-0,022197	0,012726	0,278861
LoughranMcDonald Modal Weak	0,210906	0,039714	-0,071922	0,780377
LoughranMcDonald Modal Strong	0,273110	-0,048870	0,393400	0,451457
LoughranMcDonald Litigious	0,806545	-0,157881	0,036236	0,186559
Henry Positive	0,711441	0,021944	0,046929	0,348630
Henry Negative	0,713647	0,046744	0,001093	0,416259
Tržní kapitalizace (v milionech)	0,787822	0,039511	-0,065565	-0,181263
Objem akcií	0,716777	0,157781	-0,074916	-0,079649
Typ burzy (0= NYSE, 1=Nasdaq)	-0,564150	-0,093753	-0,152457	0,089506
P/E Current	-0,042502	0,100770	0,868318	-0,003519
Price to Book Ratio	-0,079842	-0,846092	0,140794	-0,005918
Return on Equity	0,043341	-0,848394	-0,235517	-0,014760
Total Debt to Total Assets	-0,021045	0,393299	-0,467001	0,305258
Výkl.roz	5,032217	1,674073	1,244469	1,887967
Prp.celk	0,335481	0,111605	0,082965	0,125864

Varianta biquartimax normalizovaný

Proměnná	Faktor. zátěže (Biquartimax normaliz.) (DATA VŠE UPRAVENO) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)			
	Faktor (1)	Faktor (2)	Faktor (3)	Faktor (4)
LoughranMcDonald Positive	0,794361	-0,058671	0,036988	0,323730
LoughranMcDonald Uncertainty	0,605414	0,002158	0,028506	0,648801
LoughranMcDonald Negative	0,886801	0,045892	0,004812	0,280565
LoughranMcDonald Modal Weak	0,211433	-0,045338	0,120259	0,773949
LoughranMcDonald Modal Strong	0,271877	0,039434	-0,366059	0,475481
LoughranMcDonald Litigious	0,800874	0,179860	-0,027648	0,192472
Henry Positive	0,711644	-0,006174	-0,024252	0,351157
Henry Negative	0,714524	-0,030959	0,026188	0,415405
Tržní kapitalizace (v milionech)	0,788676	-0,008550	0,056449	-0,184690
Objem akcií	0,721469	-0,130674	0,074720	-0,086287
Typ burzy (0= NYSE, 1=Nasdaq)	-0,567300	0,077336	0,154525	0,081370
P/E Current	-0,037361	-0,124107	-0,864281	0,046580
Price to Book Ratio	-0,107296	0,839115	-0,161565	0,019347
Return on Equity	0,014968	0,855210	0,213514	-0,011887
Total Debt to Total Assets	-0,009286	-0,387336	0,494038	0,268719
Výkl.roz	5,038528	1,666569	1,248099	1,885530
Prp.celk	0,335902	0,111105	0,083207	0,125702

Varianta biquartimax prostý

Proměnná	Faktor. zátěže (Biquartimax pr.) (DATA VŠE UPRAVENO) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)			
	Faktor (1)	Faktor (2)	Faktor (3)	Faktor (4)
LoughranMcDonald Positive	0,849696	0,080114	-0,022993	0,108120
LoughranMcDonald Uncertainty	0,754241	0,007511	0,009695	0,468248
LoughranMcDonald Negative	0,930256	-0,021443	0,001803	0,037704
LoughranMcDonald Modal Weak	0,406124	0,042397	-0,066040	0,699374
LoughranMcDonald Modal Strong	0,386222	-0,048069	0,394191	0,358733
LoughranMcDonald Litigious	0,827782	-0,157458	0,025305	-0,030055
Henry Positive	0,778234	0,022904	0,039708	0,150417
Henry Negative	0,797413	0,048004	-0,005320	0,215757
Tržní kapitalizace (v milionech)	0,712428	0,038880	-0,080176	-0,379324
Objem akcií	0,670140	0,157515	-0,087007	-0,262968
Typ burzy (0= NYSE, 1=Nasdaq)	-0,523080	-0,093102	-0,142534	0,236035
P/E Current	-0,031308	0,099318	0,868947	-0,006106
Price to Book Ratio	-0,076305	-0,846327	0,140595	0,015687
Return on Equity	0,035664	-0,848052	-0,237713	-0,019102
Total Debt to Total Assets	0,053237	0,395095	-0,462349	0,306102
Výkl.roz	5,638243	1,675081	1,242044	1,283358
Prp.celk	0,375883	0,111672	0,082803	0,085557

Varianta equamax normalizovaný

Proměnná	Faktor. zátěže (Equamax normaliz.) (DATA VŠE UPRAVENO) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)			
	Faktor (1)	Faktor (2)	Faktor (3)	Faktor (4)
LoughranMcDonald Positive	0,841103	-0,059218	0,037744	0,168031
LoughranMcDonald Uncertainty	0,717739	0,003703	0,016949	0,522338
LoughranMcDonald Negative	0,924013	0,045143	0,009009	0,106500
LoughranMcDonald Modal Weak	0,354052	-0,042913	0,098828	0,723366
LoughranMcDonald Modal Strong	0,360502	0,043754	-0,376449	0,402448
LoughranMcDonald Litigious	0,823310	0,179111	-0,020926	0,034665
Henry Positive	0,765694	-0,005954	-0,025095	0,208380
Henry Negative	0,780331	-0,030794	0,023179	0,272669
Tržní kapitalizace (v milionech)	0,738521	-0,011713	0,073262	-0,329335
Objem akcií	0,690998	-0,133324	0,086614	-0,218782
Typ burzy (0= NYSE, 1=Nasdaq)	-0,542505	0,077952	0,144436	0,192507
P/E Current	-0,021035	-0,117565	-0,866646	0,024734
Price to Book Ratio	-0,099237	0,840591	-0,157395	0,029829
Return on Equity	0,011843	0,853545	0,220277	-0,011704
Total Debt to Total Assets	0,037638	-0,389541	0,482407	0,283768
Výkl.roz	5,533811	1,667063	1,245696	1,392157
Prp.celk	0,368921	0,111138	0,083046	0,092810

Varianta equamax prostý

Proměnná	Faktor. zátěže (Equamax pr.) (DATA VŠE UPRAVENO) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)			
	Faktor (1)	Faktor (2)	Faktor (3)	Faktor (4)
LoughranMcDonald Positive	0,855212	0,075289	0,026091	0,053695
LoughranMcDonald Uncertainty	0,782458	0,004725	-0,006402	0,419503
LoughranMcDonald Negative	0,930633	-0,027061	0,001525	-0,021203
LoughranMcDonald Modal Weak	0,449630	0,042741	0,068414	0,671971
LoughranMcDonald Modal Strong	0,409298	-0,049188	-0,392315	0,334214
LoughranMcDonald Litigious	0,823350	-0,162722	-0,022371	-0,081703
Henry Positive	0,786461	0,018667	-0,036777	0,100760
Henry Negative	0,809728	0,043933	0,008397	0,164589
Tržní kapitalizace (v milionech)	0,686885	0,032958	0,082136	-0,423953
Objem akcií	0,652736	0,152327	0,088928	-0,305687
Typ burzy (0= NYSE, 1=Nasdaq)	-0,508123	-0,088833	0,141063	0,268942
P/E Current	-0,027922	0,099092	-0,869102	-0,003544
Price to Book Ratio	-0,079675	-0,845830	-0,140472	0,024407
Return on Equity	0,028494	-0,848220	0,238179	-0,017835
Total Debt to Total Assets	0,073212	0,396203	0,462784	0,299823
Výkl.roz	5,700848	1,673310	1,241762	1,222807
Prp.celk	0,380057	0,111554	0,082784	0,081520

Varianta quartimax normalizovaný

Proměnná	Faktor. zátěže (Quartimax normaliz.) (DATA VŠE UPRAVENO) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)			
	Faktor (1)	Faktor (2)	Faktor (3)	Faktor (4)
LoughranMcDonald Positive	0,841103	-0,059218	0,037744	0,168031
LoughranMcDonald Uncertainty	0,717739	0,003703	0,016949	0,522338
LoughranMcDonald Negative	0,924013	0,045143	0,009009	0,106500
LoughranMcDonald Modal Weak	0,354052	-0,042913	0,098828	0,723366
LoughranMcDonald Modal Strong	0,360502	0,043754	-0,376449	0,402448
LoughranMcDonald Litigious	0,823310	0,179111	-0,020926	0,034665
Henry Positive	0,765694	-0,005954	-0,025095	0,208380
Henry Negative	0,780331	-0,030794	0,023179	0,272669
Tržní kapitalizace (v milionech)	0,738521	-0,011713	0,073262	-0,329335
Objem akcií	0,690998	-0,133324	0,086614	-0,218782
Typ burzy (0= NYSE, 1=Nasdaq)	-0,542505	0,077952	0,144436	0,192507
P/E Current	-0,021035	-0,117565	-0,866646	0,024734
Price to Book Ratio	-0,099237	0,840591	-0,157395	0,029829
Return on Equity	0,011843	0,853545	0,220277	-0,011704
Total Debt to Total Assets	0,037638	-0,389541	0,482407	0,283768
Výkl.roz	5,533811	1,667063	1,245696	1,392157
Prp.celk	0,368921	0,111138	0,083046	0,092810

Varianta quartimax prostý

Proměnná	Faktor. zátěže (Quartimax pr.) (DATA VŠE UPRAVENO) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000)			
	Faktor (1)	Faktor (2)	Faktor (3)	Faktor (4)
LoughranMcDonald Positive	0,855212	0,075290	0,026091	0,053695
LoughranMcDonald Uncertainty	0,782458	0,004725	-0,006402	0,419503
LoughranMcDonald Negative	0,930633	-0,027061	0,001525	-0,021203
LoughranMcDonald Modal Weak	0,449630	0,042741	0,068414	0,671971
LoughranMcDonald Modal Strong	0,409298	-0,049188	-0,392315	0,334214
LoughranMcDonald Litigious	0,823350	-0,162722	-0,022371	-0,081703
Henry Positive	0,786461	0,018667	-0,036777	0,100760
Henry Negative	0,809728	0,043933	0,008397	0,164589
Tržní kapitalizace (v milionech)	0,686885	0,032958	0,082136	-0,423953
Objem akcií	0,652736	0,152327	0,088928	-0,305687
Typ burzy (0= NYSE, 1=Nasdaq)	-0,508123	-0,088833	0,141063	0,268942
P/E Current	-0,027922	0,099092	-0,869102	-0,003544
Price to Book Ratio	-0,079675	-0,845830	-0,140472	0,024407
Return on Equity	0,028494	-0,848220	0,238180	-0,017835
Total Debt to Total Assets	0,073212	0,396203	0,462784	0,299823
Výkl.roz	5,700848	1,673310	1,241762	1,222807
Prp.celk	0,380057	0,111554	0,082784	0,081520