Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačních technologií

Bakalářská práce

Analýza sentimentu v oblasti žurnalistiky

Ilias Lotfi

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Ilias Lotfi

Informatika

Název práce

Analýza sentimentu v oblasti žurnalistiky.

Název anglicky

Sentiment analysis in the field of journalism.

Cíle práce

Bakalářská práce zkoumá problematiku vlivu emočního faktoru na objektivitu informací. Hlavním cílem práce je definovat spolehlivost a objektivitu informačních zdrojů prostřednictvím provádění analýzy sentimentu.
Dílčí cíle jsou:
- provádění analýzy sentimentu článků na jedno téma z různých informačních zdrojů.
- porovnání výsledků pro identifikaci zdroje s nejmenším emocionálním zbarvením.
- srovnání úrovně důvěry obyvatelstva v zdroje informací s úrovní jejich emočního zbarvení.
- navrhnout koncepci závislosti důvěry společnosti na úrovni emočního zbarvení textu.

Metodika

Teoretická část bakalářské práce se zaměří na výběr způsobu, jak provádět analýzu sentimentu na základě dostupných studií a shromažďování informací, která bude předmětem analýzy. Praktická část bude věnována provádění analýzy sentimentu a určení závislosti úrovně emočního zbarvení na důvěře obyvatelstva.

Doporučený rozsah práce

30–40 stran

Klíčová slova

Analýza sentimentu, informační zdroje, informační technologie, žurnalistika, Data Science

---

Doporučené zdroje informací

BBC news [online] Dostupné z: www.bbc.com
CNN [online]. Dostupné z: edition.cnn.com
New York Times [online] Dostupné z: www.nytimes.com
Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis. Springer. ISBN 9783319236544.

Předběžný termín obhajoby

2022/23 LS – PEF

Vedoucí práce

Ing. Michal Stočes, Ph.D.

Garantující pracoviště

Katedra informačních technologií

Elektronicky schváleno dne 27. 9. 2022

doc. Ing. Jiří Vaněk, Ph.D.

Vedoucí katedry

Elektronicky schváleno dne 27. 10. 2022

doc. Ing. Tomáš Šubrt, Ph.D.

Děkan

V Praze dne 29. 10. 2023

Čestné prohlášení

Prohlašuji, že svou bakalářskou práci "Analýza sentimentu v oblasti žurnalistiky" jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autor uvedené bakalářské práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 15.03.2024 _____

Poděkování

Rád bych touto cestou poděkoval Ing. Michalovi Stočesovi, Ph.D. za vedení této bakalářské práce a také kolegům ze společnosti Barclays Execution Services Limited za možnost absolvování praxe a za sdílení svých znalosti o programovacím jazyce Java.

Analýza sentimentu v oblasti žurnalistiky

Abstrakt

Tato bakalářská práce se zaměřuje na analýzu sentimentu v žurnalistice, jejímž cílem je zkoumání vlivu emočního faktoru na objektivitu vnímání informací a důvěru veřejnosti vůči médiím. Výzkum zahrnuje teoretickou část a praktickou analýzu. Je zdůrazněna důležitost porozumění vlivu emocí na vnímání zpráv.

V teoretické části práce jsou zkoumány základní principy analýzy sentimentu, metody jejího provádění a sběru dat, a také přehled existujících výzkumů. Popsán výběr použitých metod analýzy sentimentu, kde hlavním nástrojem je knihovna Stanford CoreNLP. Dále jsou popsány metody statistické analýzy pomoci programu SAS.

Praktická část práce zahrnuje vývoj programu v jazyce Java s využitím knihovny Stanford CoreNLP pro provádění analýzy sentimentu vybraných článků. Porovnávaní výsledku analýzy za účelem identifikace rozdílů v emočním zabarvení mezi médii. Dále je použit software SAS pro provádění regresní a korelační analýzy, která spojuje výsledky analýzy nálad s úrovní důvěry veřejnosti ve zkoumaná média na základě dat ze stávajících výzkumů.

Na základě provedeného výzkumu jsou formulovány závěry o vlivu emočního zabarvení na důvěru veřejnosti k zpravodajským portálům. Tato práce představuje důležitý příspěvek k pochopení dynamiky mezi médii, jejich emočním obsahem a vnímáním veřejností, což může být využito pro další výzkumy v oblasti médií a komunikace.

Klíčová slova: analýza sentimentu, emoční zabarvení, žurnalistika, metody analýzy textu, Stanford CoreNLP, analytický software SAS

Sentiment analysis in the field of journalism.

Abstract

This bachelor's thesis is devoted to the analysis of attitudes in journalism, aiming to examine the influence of the emotional factor on the objectivity of information perception and public trust towards the media. The research includes both a theoretical part and a practical analysis. The importance of understanding the impact of emotions on news perception is emphasized.

In the theoretical part of the thesis, the fundamental principles of sentiment analysis, methods of its execution and data collection, as well as an overview of existing research are examined. This section also describes the chosen methods of sentiment analysis, where the main tool is the Stanford CoreNLP library. Methods of statistical analysis using the SAS program are also described.

The practical part of the work involves the development of a program in Java using the Stanford CoreNLP library to conduct sentiment analysis of selected articles. The results of the analysis are then compared to identify differences in emotional tone between the media. Subsequently, SAS software is used for regression and correlation analysis, linking the results of the sentiment analysis with the level of public trust in the media under study, based on data from existing research.

Based on the conducted research, conclusions are drawn about the impact of emotional tone on public trust towards the media. This work represents an important contribution to understanding the dynamics between the media, their emotional content, and public perception, which can be used for further research in the field of media and communication.

Keywords: sentiment analysis, emotional tone, journalism, text analysis methods, Stanford CoreNLP, SAS analytical software

Obsah

1 Úvod

Moderní informační pole se vyznačuje obrovským množstvím dostupných informací, které získávají prostřednictvím médií: zprávy, analytické články, komentáře atd. pokrývají širokou škálu událostí a témat. Nicméně, nesou v sobě nejen objektivní informace, ale i emoční zabarvení, které může ovlivnit vnímání a hodnocení prezentovaných událostí. Význam tohoto aspektu vzrůstá v kontextu digitální éry, kdy se informace šíří okamžitě.

Tato bakalářská práce je zaměřena na analýzu sentimentu v žurnalistice, s důrazem na studium vztahu mezi emočním zabarvením textu a jeho vnímáním publikem. Tento přístup umožní hlouběji porozumět mechanismům vlivu médií na veřejné vědomí a určit, jak emoce ovlivňují důvěru k zpravodajským portálům.

2 Cíl práce a metodika

2.1 Cíl práce

Hlavním cílem této bakalářské práce je posoudit vliv emočního zabarvení článků na důvěru veřejností k zpravodajským portálům. Pro dosažení tohoto cíle práce zahrnuje několik dílčích cílů:

- Provést analýzu sentimentu v článcích různých médií na stejné téma. K tomu je třeba zvolit metodu analýzy sentimentu, shromáždit vzorek článků a provést analýzu emočního zabarvení.
- Porovnat výsledky analýzy sentimentu mezi různými zpravodajskými portály. To umožní určit rozdíly v přístupech k prezentaci informací.
- Porovnat úroveň důvěry veřejnosti v média s úrovní emočního zabarvení textů. K tomu budou použita data z předchozích výzkumů a otevřených zdrojů.
- Navrhnout koncept závislosti důvěry veřejnosti k zpravodajským portálům na úrovni emočního zabarvení článků.

2.2 Metodika

Metodika práce kombinuje teoretické a praktické přístupy. V teoretické části bude proveden přehled stávající literatury a výzkumů týkajících se analýzy sentimentu v médiích a jeho vlivu na veřejné vnímání. Praktická část zahrnuje vytvoření programu v jazyce Java s využitím knihovny Stanford CoreNLP pro provádění analýzy sentimentu vybraných článků. Dále bude použit software SAS k provádění statistické analýzy dat a k identifikaci závislosti mezi emočním zabarvením článků a úrovní důvěry vůči médiím

Tato práce se tak poskytuje komplexní pochopení vlivu emocí v článcích na důvěru k informačním zdrojům.

3 Teoretická východiska

3.1 Historie analýzy sentimentu

Analýza sentimentu vznikla jako oblast výzkumu na rozhraní informatiky, lingvistiky a umělé inteligence. Tento přístup je zaměřen na identifikaci a studium emočního zabarvení textových dat, což je možné s rozvojem technologií zpracování přirozeného jazyka a strojového učení. (Erik Cambria, Amir Hussain, 2016).

První práce v této oblasti se objevily na počátku 21.století, kdy vědci začali používat algoritmy strojového učení k analýze emocí v textech. Rozvoj internetu a sociálních médií vedl ke zvýšení množství textových informací dostupných pro analýzu, což podnítilo zájem o průzkum nálad (Bing Liu, 2012).

3.2 Vývoj analýzy sentimentu

Základní teorie a přístupy v analýze sentimentu se postupem času vyvíjely. V raných pracích byla pozornost zaměřena především na klasifikaci celého textů do pozitivních, negativních a neutrálních kategorií. Postupem času se přístupy staly složitějšími a mnohostrannými, včetně analýzy sentimentu na úrovni jednotlivých vět a dokonce i slov. V posledních letech byl vyvinut přístup koncepční úrovně analýzy sentimentu, který umožňuje přesnější hodnocení emocionálních výrazů na základě obecného kontextu a sémantických vazeb v textu (Erik Cambria, Amir Hussain, 2016).

3.2.1 Aplikace analýzy sentimentu v oblasti žurnalistiky

Aplikace analýzy nálad v oblasti žurnalistiky se stala relevantní s rostoucí rolí médií ve veřejnosti. Emoční zabarvení zpravodajských materiálů má významný dopad na veřejné mínění, takže je důležité porozumět tomu, jak různé emoční prvky ovlivňují vnímání informací. Analýza sentimentu tedy poskytuje nástroje pro studium této dynamiky, což umožňuje posoudit nejen objektivní obsah zpráv, ale také jejich emočního kontext (D'Andrea A., Ferri F., Grifoni P., 2015).

V moderních studiích se analýza sentimentu používá ke studiu různých aspektů interakce mezi médii a publikem, včetně otázek důvěry, emočního zapojení a dopadu na utváření veřejného mínění. Analýza nálad tak pomáhá identifikovat a studovat skryté vzorce

v prezentaci zpráv, což je důležité pro pochopení mechanismů dopadu médií na veřejné vědomí (Bing Liu, 2012).

3.3 Metody analýzy sentimentů

Analýza sentimentu, vyvíjející se v rámci interdisciplinárního výzkumu, zahrnuje různé techniky, které odrážejí vývoj této oblasti.

- Lexikální přístup: původní metody analýzy sentimentu byly založeny na lexikálních základech, kde každému slovo byla přiřazena určitá emoční hodnota. Tyto metody jsou snadno implementovatelné, ale mají omezení související s kontextovými rysy jazyka (Erik Cambria, Amir Hussain, 2016).
- Strojové učení: s rozvojem strojového učení se objevil složitější přístup založený na klasifikaci textů podle nálady pomocí předem vyškolených modelů. Tento přístup umožňuje zohlednit širší kontext a nuance jazyka (Bing Liu, 2012).
- Hybridní metody: kombinace lexikálních a strojově učených přístupů umožňuje dosáhnout vyšší přesnosti a hloubky analýzy vzhledem ke statistické významnosti slov a jejich sémantické vazbě v kontextu (Erik Cambria, Amir Hussain, 2016).
- Hluboké učení: v popředí metodiky analýzy nálad jsou metody hlubokého učení, jako jsou konvoluční a rekurentní neuronové sítě. Jsou schopni zpracovat velké množství dat a identifikovat složité emocionální vzory, díky čemuž jsou ideální pro komplexní analýzu textů (D'Andrea A., Ferri F., Grifoni P., 2015).

Výběr konkrétní metody pro analýzu nálad závisí na specifikách zkoumaných dat a zadaných úkolech. V kontextu žurnalistiky, kde je důležité porozumět vlivu emočního zabarvení na vnímání publika, lze použít některý z těchto přístupů v závislosti na cílech studie.

Tyto metody poskytují základ pro praktickou část práce, která provede analýzu sentimentu článků z různých médií s cílem posoudit a porovnat jejich emocionální dopad a úroveň důvěry veřejnosti. Použití těchto metod v kontextu žurnalistiky může poskytnout cenné znalosti o tom, jak zpravodajský obsah stanoví veřejné mínění a ovlivňuje důvěru v informační zdroje.

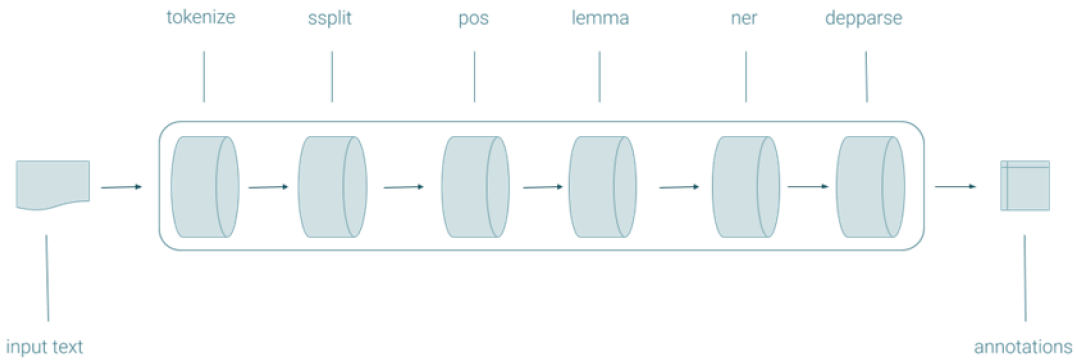
3.3.1 Stanford Core NLP

V kontextu tohoto výzkumu, jehož cílem je hodnotit emoční zabarvení článků z různých médií a jeho vliv na důvěru veřejnosti, se použití Stanford CoreNLP jeví jako nejvhodnější. Tento nástroj umožňuje efektivně zpracovávat a analyzovat velké objemy textu, čímž zajišťuje potřebnou kvalitu analýzy sentimentu. (Stanford NLP Group, 2020)

Stanford CoreNLP je integrovaný nástroj pro zpracování přirozeného jazyka, který poskytuje široký spektrum analytických funkcí. Základní Pipeline knihovny Stanford CoreNLP organizuje proces zpracování textu jako sérii kroků, přičemž každý krok vykonává určitou úlohu analýzy textu. Pipline přijímá nezpracovaný text, spouští řadu anotátorů NLP na text a generuje konečnou sadu anotací. Níže je popsán význam každého z uvedených kroků:

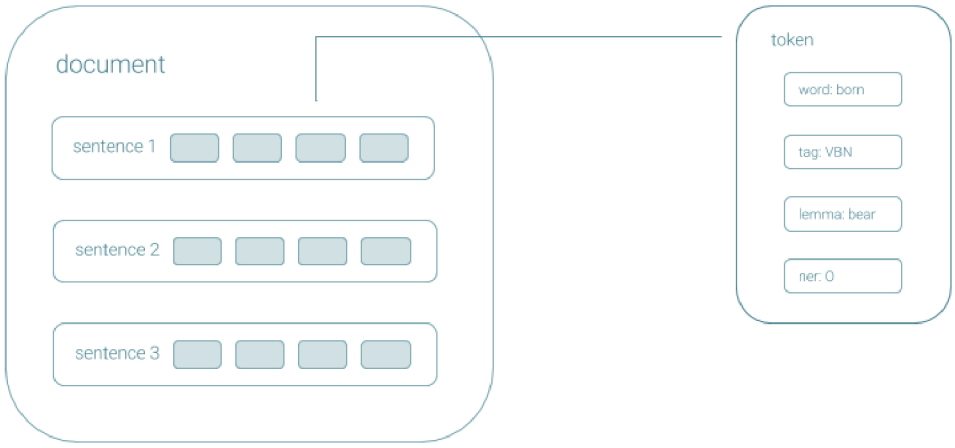
- Input text: Počáteční fáze, ve které je do Pipeline vstupován zdrojový text pro zpracování. Jedná se o text, který uživatel chce analyzovat, a může zahrnovat věty, odstavce nebo dokumenty. (Stanford NLP Group, 2020)
- Tokenize: Na tomto kroku je text rozdělen na tokeny, které obvykle představují slova a interpunkční znaménka. Tokenizace je základním krokem, protože většina následujících kroků analýzy pracuje s jednotlivými tokeny. (Stanford NLP Group, 2020)
- Ssplit (Sentence Splitting): Po tokenizaci je text rozdělen na věty. Rozdělení na věty je důležité, protože mnoho úloh a modelů NLP předpokládá práci s textem na úrovni vět, je třeba přesně určit hranice vět pro správnou funkci následujících etap. (Stanford NLP Group, 2020)
- POS (Part of Speech tagging): Na tomto kroku je každému tokenu přiřazena značka části řeči (podstatné jméno, sloveso, přídavné jméno atd.). Analýza částí řeči pomáhá v pochopení gramatické struktury vět a je důležitá pro mnohé následující úkoly, jako je syntaktický rozbor a rozpoznávání entit. (Stanford NLP Group, 2020)
- Lemma (Lemmatizace): Lemmatizace převádí slova na jejich kanonickou formu (lemmu), což pomáhá snížit morfologickou diverzitu slov a zjednodušuje jejich analýzu. Například slova “run”, “runs”, “ran” budou převedena na základní formu “run”. (Stanford NLP Group, 2020)

- NER (Named Entity Recognition): Na tomto kroku jsou identifikovány a klasifikovány pojmenované entity (jména osob, organizací, geografických názvů, dat atd.). To umožňuje extrahovat z textu konkrétní informace o entitách a jejich typech. (Stanford NLP Group, 2020)
- Depparse (Dependency Parsing): Úkolem je vytvoření stromu závislostí pro věty, kde uzly jsou slova a hrany představují gramatické vztahy mezi nimi. To umožňuje pochopit strukturu věty a vztahy mezi slovy. (Stanford NLP Group, 2020)
- Anotace: Na posledním kroku jsou informace získané v předchozích krocích agregovány do anotací. Anotace mohou zahrnovat různé údaje o zpracovaném textu, jako jsou značky částí řeči, lemmata, entity a struktura závislostí. Tyto anotace činí výsledky zpracování dostupnými pro uživatele nebo pro následné procesy analýzy. V případě tohoto výzkumu výsledkem bude bodové hodnoceni v rozsahu od 0 do 4, kde 0 – Very Negative, 4 – Very Positive (Stanford NLP Group, 2020)



Obrázek 1: Pipeline Stanford CoreNLP. Zdroj: Stanford NLP group, 2020

Pipeliny produkují CoreDocuments, datové objekty, které obsahují veškeré informace o anotacích. (Stanford NLP Group, 2020)



Obrázek 2: Core document Stanford CoreNLP. Zdroj: Stanford NLP group, 2020

Stanford CoreNLP poskytuje tyto a další funkce prostřednictvím pohodlného API, které lze integrovat do různých aplikací pro zpracování přirozeného jazyka. Důležitou vlastností je jeho modularita: uživatelé mohou vybrat pouze ty komponenty, které potřebují pro svůj úkol, což činí Stanford CoreNLP flexibilním a mocným nástrojem pro výzkumníky a vývojáře v oblasti NLP.

3.4 Statistická analýza

Statistická analýza představuje komplex metod zpracování dat, který umožňuje interpretovat shromážděné informační soubory, identifikovat v nich vzorce, trendy a anomálie. Zahrnuje popisnou statistiku, která popisuje základní charakteristiky datové sady, a inferenční statistiku, umožňující dělat zobecnění na základě výběru dat o větší populaci. Statistická analýza je široce využívána v různých oblastech vědy a praxe pro ověřování hypotéz a modelů, jakož i pro podporu přijímání odůvodněných rozhodnutí. (Abdishakur Hassan, 2022)

3.4.1 Program SAS

SAS (Statistical Analysis System) je komplexní software pro statistickou analýzu dat vyvinutý společností SAS Institute Inc. Poskytuje uživatelům nástroje pro provádění různých statistických analýz, správu dat, obchodní analýzu a grafické znázornění výsledků.

SAS je široce používán v průmyslu, výzkumu a akademických kruzích díky své kapacitě, flexibilitě a široké škále možností. (SAS, 2024)

SAS byl vyvinut v roce 1976 společností SAS Institute Inc., kterou založili James Goodnight, John Sall, Anthony Barr a Jane Helwig. Původně byl SAS používán pro analýzu zemědělských dat, ale postupem času se jeho možnosti výrazně rozšířily. (SAS, 2024)

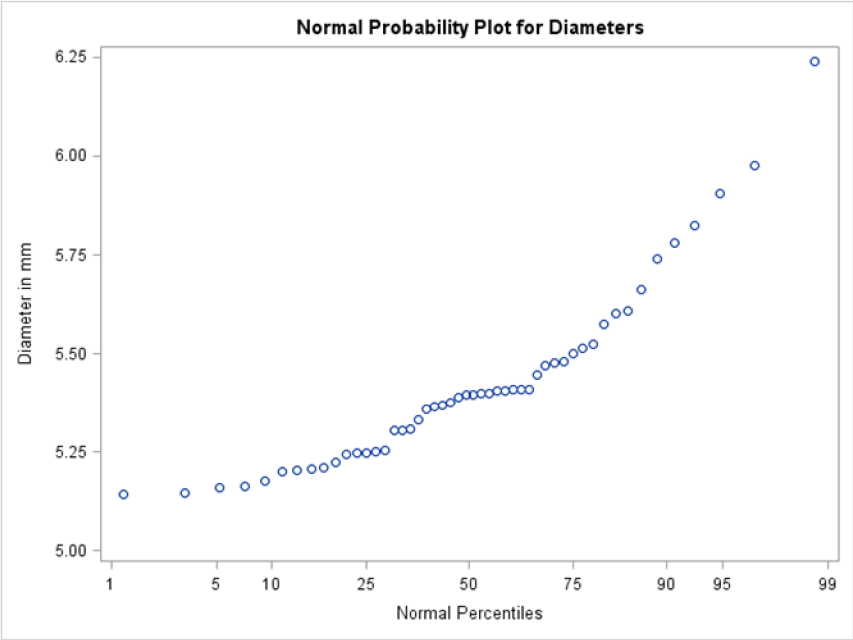
3.4.2 Vizualizace dat pomoci programu SAS

SAS používá k vizualizaci dat a analýze výsledků mnoho různých typů grafů. Každý z nich pomáhá výzkumníkům lépe porozumět distribuci, struktuře a možným vztahům v datech. Zde je popis některých klíčových typů grafů, které lze vytvořit v SAS:

3.4.2.1 Normal Probability Plot

Tento graf se používá k posouzení, jak dobře soubor dat odpovídá normálnímu rozložení. Porovnává pozorované hodnoty s teoretickými kvantilami normálního rozdělení. (SAS Institute Inc, 2024)

Velmi užitečné ve statistické analýze pro kontrolu předpokladů normality, což je klíčový požadavek pro mnoho parametrických statistických testů.

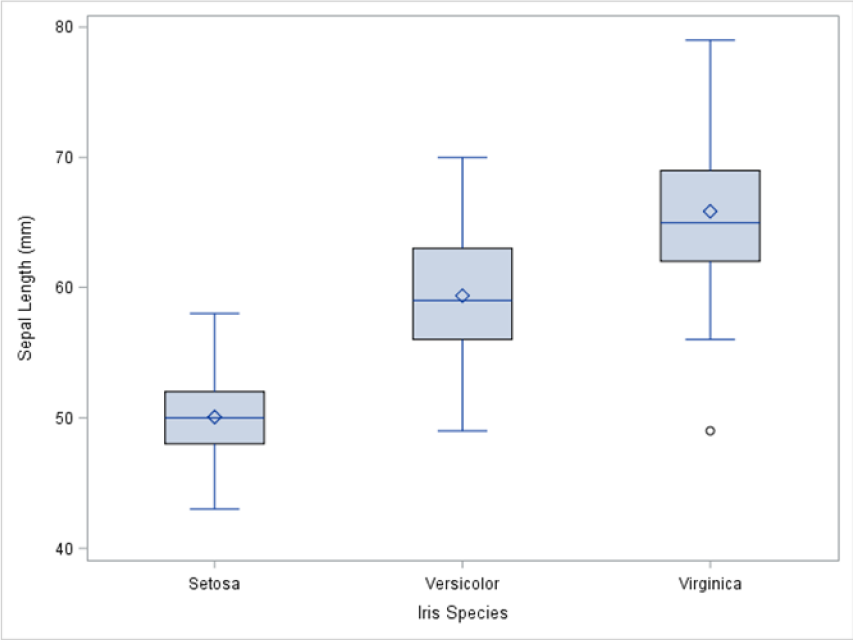


Obrázek 3: Normal Probability Plot. Zdroj: SAS Institute Inc, 2024

3.4.2.2 Boxplot

Krabicový diagram zobrazuje rozdělení dat prostřednictvím kvartilů a zvýrazňuje možné výstřelky. Centrální linie krabice reprezentuje medián, okraje krabice první a třetí kvartily, zatímco "vousy" ukazují rozptyl dat. (SASnrd, 2024)

Krabicový diagram je ideální pro srovnání rozdělení mezi několika skupinami a identifikaci výstřelků. (SASnrd, 2024)

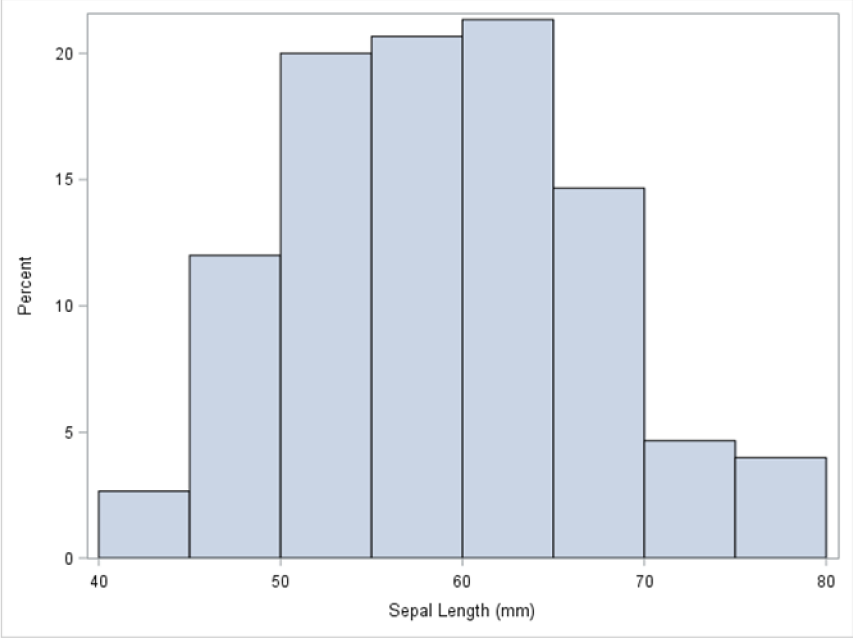


Obrázek 4: Boxplot. Zdroj: SASnrd, 2024

3.4.2.3 Histogram

Histogram ukazuje rozložení frekvence nebo pravděpodobnosti hodnot v sadě dat. Každý sloupec histogramu představuje rozsah hodnot a jeho výška odpovídá frekvenci (nebo pravděpodobnosti) hodnot v tomto rozsahu. (SASnrd, 2024)

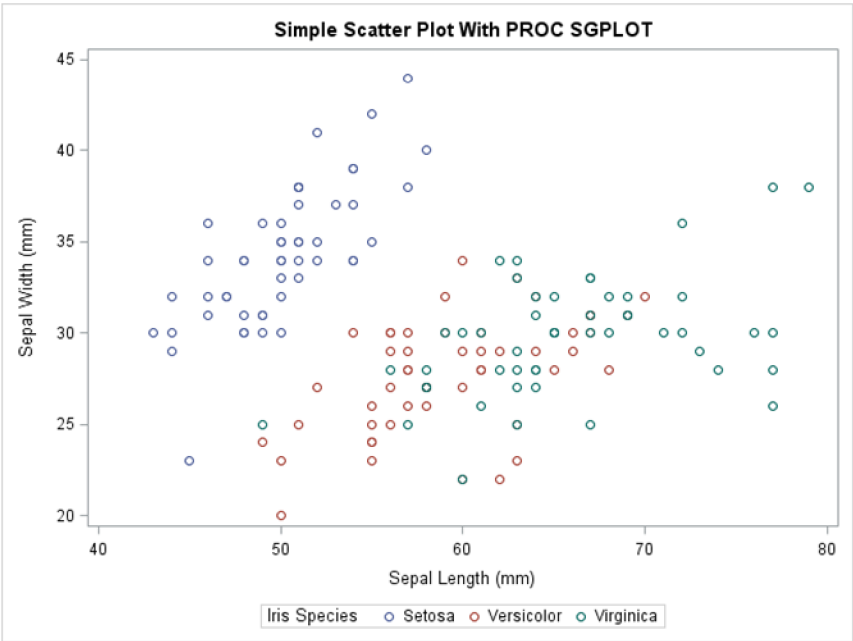
Histogramy se používají k vizualizaci distribuce dat, což pomáhá určit, zda je distribuce normální, zkosená atd. (SASnrd, 2024)



Obrázek 5: Histogram. Zdroj: SASnrd, 2024

3.4.2.4 Korelační diagram (Scatter Plot)

Korelační diagram zobrazuje hodnoty dvou proměnných jako body na dvourozměrné rovině, kde každá osa reprezentuje jednu z proměnných. To umožňuje vizualizovat vztah mezi proměnnými. Používá se k identifikaci korelací, trendů a možných anomálií mezi páry proměnných. (SASnrd, 2024)



Obrázek 6: Korelační diagram. Zdroj: SASnrd, 2024

3.4.3 Test normality

Test normality je statistický procedurální test, který kontroluje, zda se distribuce dat řídí normálním (gaussovským) rozdělením. Tento test je důležitým krokem v mnoha statistických analýzách, protože mnoho statistických metod vyžaduje, aby byla data distribuována normálně, aby byly výsledky přesné a spolehlivé. Mezi takové metody patří parametrické testy, včetně t-testů, ANOVA (disperzní analýza), regresní analýzy a další. (Cody, R., Smith, J. K., 2005)

V případech, kdy data nedodržují normální distribuci, mohou být použity neparametrické metody. Tyto metody nepředpokládají konkrétní formu distribuce dat, a proto jsou vhodné pro analýzu dat s jakýmkoli typem distribuce. Neparametrické testy, jako jsou Mann-Whitney test, Wilcoxonův test, Craskel-Wallis test a další, poskytují alternativní způsoby, jak otestovat hypotézy o medianech, distribuci a vztahu mezi proměnnými bez přísných požadavků na formu distribuce dat. (Cody, R., Smith, J. K., 2005)
Existuje několik různých testů normality, včetně:

- **Shapiro-Wilkovo kritérium**: jeden z nejsilnějších testů normality pro malé datové vzorky.
- **Kolmogorov-Smirnovovo kritérium**: často se používá pro velké vzorky, může porovnat vzorek s jakoukoli distribucí.
- **Andersonovo-Darlingovo kritérium**: další test, který je obzvláště citlivý na distribuční ocasy.

Správný výběr statistických metod je založen na předpokladu distribuce dat. Nesprávný předpoklad může vést k nesprávným závěrům. V reálných datech je ideální normalita vzácná a malé odchylky od normality nemusí mít v některých případech významný vliv na výsledky analýzy. V takových situacích lze použít metody transformace dat nebo výběr neparametrických analytických metod. (DATAtab Team, 2024)

3.4.4 Regresní a Korelační analýza

Regresní a korelační analýzy jsou dvě důležité statistické metody používané ke studiu vztahů mezi proměnnými. Obě metody umožňují analyzovat vazby mezi dvěma nebo více proměnnými, ale slouží různým účelům a jsou založeny na různých principech. (Elliott, A. C., & Woodward, W. A., 2015)

3.4.4.1 Regresní analýza

Regresní analýza se používá k modelování vztahu mezi závislou proměnnou (nebo předpovídanou proměnnou) a jednou nebo více nezávislými proměnnými (nebo prediktory). Cílem regresní analýzy je předpovědět hodnotu závislé proměnné na základě hodnot nezávislých proměnných. (Elliott, A. C., & Woodward, W. A., 2015)

Příklady modelů:

- **Lineární regrese:** modeluje lineární vztah mezi závislou a nezávislou proměnnou.
- **Vícenásobná regrese:** používá několik nezávislých proměnných k předpovědi hodnoty závislé proměnné.
- **Logistická regrese:** používá se k modelování pravděpodobnosti výskytu nějaké události (ano/ne).

3.4.4.2 Korelační analýza

Korelační analýza měří stupeň a směr lineárního vztahu mezi dvěma proměnnými. Výsledkem korelační analýzy je korelační koeficient, který se může pohybovat od -1 do 1, Kde -1 znamená dokonalou zpětnou korelaci, 0 je nepřítomnost korelace a 1 je ideální přímá korelace. (Elliott, A. C., & Woodward, W. A., 2015)

Příklad:

- **Pearsonův korelační koeficient:** měří lineární závislost mezi dvěma proměnnými.
- **Spearmanův korelační koeficient:** používá se k měření vztahu mezi hodnostmi dvou proměnných.
- **Korelační koeficient Kendall**: další metoda měření rankové korelace.

Regresní analýza se snaží předpovědět hodnotu jedné proměnné na základě druhé, zatímco korelační analýza měří sílu a směr vazby mezi proměnnými.

4 Vlastní práce

4.1 Sběr dat

První fází praktické části této bakalářské práce je sběr dat, která bude tvořit základ pro následnou analýzu sentimentu.

4.1.1 Použiti existujícího studia

Na základě cíle výzkumu bylo rozhodnuto použít data portálu yougov.com ze studia „Trust in Media 2023“, což je hodnocení zpravodajských publikací na základě úrovně důvěry veřejnosti. Na základě studia bylo vybráno 20 zpravodajských portálů, které reprezentují rozmanitost mediálního prostředí a mají různou úroveň důvěry mezi diváky. To umožňuje provádět komplexní analýzu sentimentu zahrnující širokou škálu názorů a úhlů pohledu. (Linley S., 2023)

Níže představeno zobrazení čistého skóre důvěry mezi dospělými občany USA: rozdíl v procentních bodech mezi procenty dospělých občanů USA, kteří uvádějí, že každý zpravodajský portál je důvěryhodný nebo velmi důvěryhodný, a procenty těch, kteří tvrdí, že je nedůvěryhodný nebo velmi nedůvěryhodný

Portal	Rating
PBS	30
The BBC	29
The Wall Street Journal	24
Forbes	23
The Associated Press news	22
ABC News	21
USA Today	21
CBS	20
Reuters	20
NBC	19
National Public Radio news	16
The Guardian	15
Bloomberg	10
CNN	7
Yahoo News	7
Fox News	3
The Daily Beast	1
Daily Kos	-1
The Daily Caller	-4
Infowars	-16

Tabulka 1: Úrovně důvěry veřejnosti k zpravodajským portálům. Zdroj: Linley S., 2023

4.1.2 Příprava dat k analýze

Z každého vybraného zpravodajského portálu bylo náhodně vybráno 5 článků na téma Umělá Inteligence publikovaných v listopadu 2023. Volba tohoto konkrétního tématu je důsledkem jeho relevance a potenciálního vlivu na veřejné mínění, což jej činí zvláště zajímavým pro analýzu sentimentu.

Shromážděné materiály byly podrobeny pečlivému předzpracování s cílem odstranit neinformativní prvky, jako jsou odkazy, reklamní bloky a další rušivé prvky, které by mohly ovlivnit výsledky analýzy.

Tato fáze položila základ pro další analýzu sentimentu v kontextu žurnalistiky.

4.2 Analýza Sentimentu pomoci Stanford CoreNLP

4.2.1 Práce se závislostmi

Analýza sentimentu je prováděna pomoci knihovny Stanford CoreNLP verzi 4.2.2 integrované do aplikace v jazyce Java.

Pro získání přístupu k funkcím knihovny byly do projektu implementovány dvě závislosti pomocí Apache Maven – nástroje pro správu, řízení a automatizaci buildů aplikací.

4.2.1.1 Základní závislost

První závislost obsahuje pouze základní knihovnu Stanford CoreNLP bez předem načtených modelů. Poskytuje základní nástroje a funkce pro zpracování přirozeného jazyka, jako jsou tokenizace, rozdělení na věty, lemmatizace atd.

```
<dependency>
    <groupId>edu.stanford.nlp</groupId>
    <artifactId>stanford-corenlp</artifactId>
    <version>4.2.2</version>
</dependency>
```

Obrázek 7: Fragment kódu. Základní závislost Stanford CoreNLP. Zdroj: vlastní zpracování

4.2.1.2 Pomocná závislost

Druhá závislost s klasifikátorem models obsahuje trénované modely potřebné pro provádění různých úkolů textové analýzy pomocí Stanford CoreNLP. Modely obsahují data a parametry, které jsou knihovnou využívány pro analýzu přirozeného jazyka. Bez těchto

modelů knihovna nebude schopna správně fungovat, jelikož jí budou chybět data potřebná pro provádění nezbytných textových analýz.

```
<dependency>
    <groupId>edu.stanford.nlp</groupId>
    <artifactId>stanford-corenlp</artifactId>
    <version>4.2.2</version>
    <classifier>models</classifier>
</dependency>
```

Obrázek 8: Fragment kódu. Pomocná závislost models Stanford CoreNLP. Zdroj: vlastní zpracování

4.2.2 Provádění analýzy sentimentu

Pro provedení analýzy sentimentu byla vytvořena konzolová aplikace, která využívá texty předem vybraných článků jako vstup. Aplikace spouští sérii anotátorů na daném textu a provádí analýzu sentimentu. Výsledkem této analýzy je bodové hodnocení úrovně emočního zabarvení textu na škále od 0 do 4.

Dále jsou popsány klíčové fragmenty kódu, které zajišťují analýzu sentimentu.

4.2.2.1 Nastavení prostředí a vytvoření pipeline

Prvním krokem v procesu analýzy je konfigurace prostředí pro zpracování textu. Pro tento účel je vytvořena instance třídy Properties, do které jsou vložena nastavení určující knihovně, které komponenty analýzy mají být použity.

```
// Nastavení pro zpracování textu
Properties props = new Properties();
props.setProperty("annotators", "tokenize, ssplit, pos, lemma, ner, parse, sentiment");
props.setProperty("tokenize.language", "en");
```

Obrázek 9: Fragment kódu. Nastavení pro zpracování textu. Zdroj: vlastní zpracování

Zde je „props“ nastaveno na použití řady anotátorů:

- tokenizace (tokenize),
- rozdělení na věty (ssplit),
- určení částí řeči (pos),
- lemmatizace (lemma),
- rozpoznávání pojmenovaných entit (ner),
- syntaktická analýza (parse) a analýza sentimentu (sentiment).
- jazyk zpracovávaného textu – angličtina (en).

Dalším krokem je vytvoření StanfordCoreNLP pipeline, která integruje všechny uvedené anotátory a je připravená na zpracování textových dat.

```
// Vytvoření pipeline s uvedenými nastaveními
StanfordCoreNLP pipeline = new StanfordCoreNLP(props);
```

Obrázek 10: Fragment kódu. Vytvoření pipeline s uvedenými nastaveními. Zdroj: vlastní zpracování

4.2.2.2 Analýza sentimentu

Text určený k analýze sentimentu umístěn do proměnné text, poté je vytvořen objekt Annotation s tímto textem pro následné zpracování všemi anotátory v pipline.

```
// Text k zpracování
String text = "Some text to analyze";

// Vytvoření prázdné Annotation s naším textem
Annotation document = new Annotation(text);
```

Obrázek 11: Fragment kódu. Text k zpracování a vytvoření Annotation k dokumentu. Zdroj: vlastní zpracování

Hlavní částí aplikace je analýza sentimentu každé věty v textu. Pro tento účel jsou z objektu document extrahovány věty a následně je pro každou větu určen sentiment.

```
// Analýza sentimentu každé věty
List<Integer> sentiments = new ArrayList<>();
List<CoreMap> sentences = document.get(CoreAnnotations.SentencesAnnotation.class);
for (CoreMap sentence : sentences) {
    String sentiment = sentence.get(SentimentCoreAnnotations.SentimentClass.class);
    int sentimentValue = convertSentiment(sentiment);
    sentiments.add(sentimentValue);
}
```

Obrázek 12: Fragment kódu. Analýza sentimentu každé věty. Zdroj: vlastní zpracování

Každá věta je analyzována a její sentiment je převeden na číselnou hodnotu pomocí pomocné funkce convertSentiment.

```java
// Metoda pro převod sentimentu na číselnou hodnotu
1 usage
private static int convertSentiment(String sentiment) {
    switch (sentiment) {
        case "Very negative":
            return 0;
        case "Negative":
            return 1;
        case "Neutral":
            return 2;
        case "Positive":
            return 3;
        case "Very positive":
            return 4;
        default:
            return 2; //Default Je považováno za neutrální
    }
}
```

Obrázek 13: Fragment kódu. Metoda pro převod sentimentu na číselnou hodnotu. Zdroj: vlastní zpracování

4.2.2.3 Agregace výstupu

Dále jsou shromážděné hodnoty agregovány k získání celkového ukazatele sentimentu textu. Výsledek je vypsán na konzoli spolu s odpovídajícím textovým popisem.

```java
// Agregace výsledků
double totalSentiment = 0;
for (Integer sentiment : sentiments) {
    totalSentiment += sentiment;
}

// Určení celkového sentimentu textu
double averageSentiment = totalSentiment / sentiments.size();
System.out.println("Celkový sentiment textu: ("+averageSentiment+")");
```

Obrázek 14: Fragment kódu. Agregace výsledků a určení celkového sentimentu textu. Zdroj: vlastní zpracování

4.2.3 Výsledky analýzy sentimentu

Výsledkem práce programu jsou číselné hodnoty pro každý z analyzovaných článků. Na základě těchto hodnot byla vypočítána průměrná hodnota emočního zabarvení pro každý zpravodajský portál. Získané hodnoty byly porovnány s odpovídajícími hodnotami ratingu důvěry veřejnosti a strukturovány do tabulky.

Portal	Sentiment článků					Sentiment průměr	Rating
	1	2	3	4	5		
PBS	1.780488	1.8125	1.851852	1.666667	1.925	1.807301265	30
The bbc	1.333333	1.787234	1.7	1.704545	1.8	1.665022566	29
The Wall Street Journal	1.526316	1.666667	1.75	1.8	2	1.748596491	24
Forbes	2.275862	2.070588	2.133333	2.2	2.159091	2.167774909	23
The Associated Press news	1.7	1.775	1.888889	1.910448	1.787879	1.812443088	22
ABC News	2.153846	1.823529	1.826087	1.45	1.923077	1.835307889	21
USA Today	1.95	2.176471	1.829787	1.615385	1.672414	1.848811246	21
CBS	1.85	1.648649	1.882353	2.117647	1.545455	1.808820639	20
Reuters	1.714286	2.022222	2.2	1.833333	1.5	1.853968254	20
NBC	2	1.970588	2.111111	1.62963	1.842105	1.910686848	19
National Public Radio news	1.771429	1.583333	1.738095	1.727273	1.881356	1.74029716	16
The Guardian	1.716667	2.068966	1.932584	1.9375	1.888889	1.908921068	15
Bloomberg	1.636364	1.97561	1.949367	2	2	1.912268096	10
CNN	1.88	2.106061	1.955556	2	1.695652	1.927453667	7
Yahoo News	2.1	2.1875	1.5	1.75	1.882353	1.883970588	7
Fox News	1.607143	1.533333	1.6	2.071429	1.592593	1.680899471	3
The Daily Beast	1.780488	2.2	1.681818	2	1.636364	1.859733925	1
Daily Kos	1.466667	1.681818	1.791667	1.73913	2.270833	1.790023057	-1
The Daily Caller	1.571429	1.71875	1.823529	1.5	1.636364	1.650014324	-4
Infowars	1.928571	2	1.769231	1.388889	2.214286	1.86019536	-16

Obrázek 15: Výsledky analýzy sentimentu. Zdroj: vlastní zpracování

Výsledky ukazují, že úroveň emočního zabarvení jednotlivých článků se pohybuje od 1.33 (negativní zabarvení s malým odchýlením směrem k neutrálnímu) až po 2.275862 (neutrální zabarvení s malým odchýlením směrem k pozitivnímu). Průměrná hodnota pro každý zpravodajský portál se nachází v rozmezí od 1.65 do 1.92 (negativní zabarvení s odchýlením směrem k neutrálnímu), s výjimkou Forbes - 2.167774909 (neutrální zabarvení s malým odchýlením směrem k pozitivnímu).

4.3 Statistická analýza pomocí programu SAS

Tato část provádí statistickou analýzu dat, jejímž cílem je prozkoumat možné vztahy mezi hodnocením a úrovní nálady publikací na různých informačních portálech. Předpokládá se, že tyto dvě proměnné mohou být vzájemně propojeny, protože celkový tón nebo nálada publikací může ovlivnit jejich popularitu a tím i hodnocení portálu.

K dosažení stanoveného cíle byly splněny následující úkoly:

1. Provedení testu normality rozdělení proměnných pro zjištění, zda data splňují předpoklady vybraných statistických metod.
2. Analýza korelace mezi hodnocením portálů a úrovní sentimentu jejich publikací k identifikaci přítomnosti a povahy vztahu mezi těmito proměnnými.
3. Regresní analýza k posouzení vlivu úrovně nálady na portálové hodnocení, která prohloubí pochopení dynamiky interakce mezi analyzovanými proměnnými.

4.3.1 Test normality

Pro provedeni testu normality se používá procedura univariate:

```
proc univariate data= mojetabulka normal plot;
histogram duver /normal;
qqplot duver /normal (mu=est sigma=est);
var duver;
run;

proc univariate data= mojetabulka normal plot;
histogram zabarveni /normal;
qqplot zabarveni /normal (mu=est sigma=est);
var zabarveni;
run;
```

Obrázek 16: SAS. Procedura pro test normality. Zdroj: vlastní zpracování

Ten příkaz poskytuje testy Shapiro-Wilka, Kolmogorova-Smirnova a další testy normality. Kromě toho při testování se vykreslí několik typů grafů, které pomohou vizuálně posoudit, zda data odpovídají normálnímu rozdělení. Mezi tyto grafy patří: Histogram, Q-Q plot a Box plot.

4.3.1.1 Test normality pro důvěr

K zjištění normality proměnné důvěr se používá Shapiro-Wilkův test. P-hodnota testu je 0.1724 a je větší než 0.05, tím pádem se nulová hypotéza přijímá a proměnná důvěr má normální rozdělení. Kolmogorov-Smirnov, Cramer-von Mises a Anderson-Darling rovněž neposkytují důvody pro odmítnutí hypotézy o normalitě (p-hodnoty 0.0877, 0.1088 a 0.1433 odpovídající).

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.932489	Pr < W	0.1724
Kolmogorov-Smirnov	D	0.180068	Pr > D	0.0877
Cramer-von Mises	W-Sq	0.099239	Pr > W-Sq	0.1088
Anderson-Darling	A-Sq	0.546853	Pr > A-Sq	0.1433

Obrázek 17: SAS. Test normality pro duver. Zdroj: vlastní zpracování

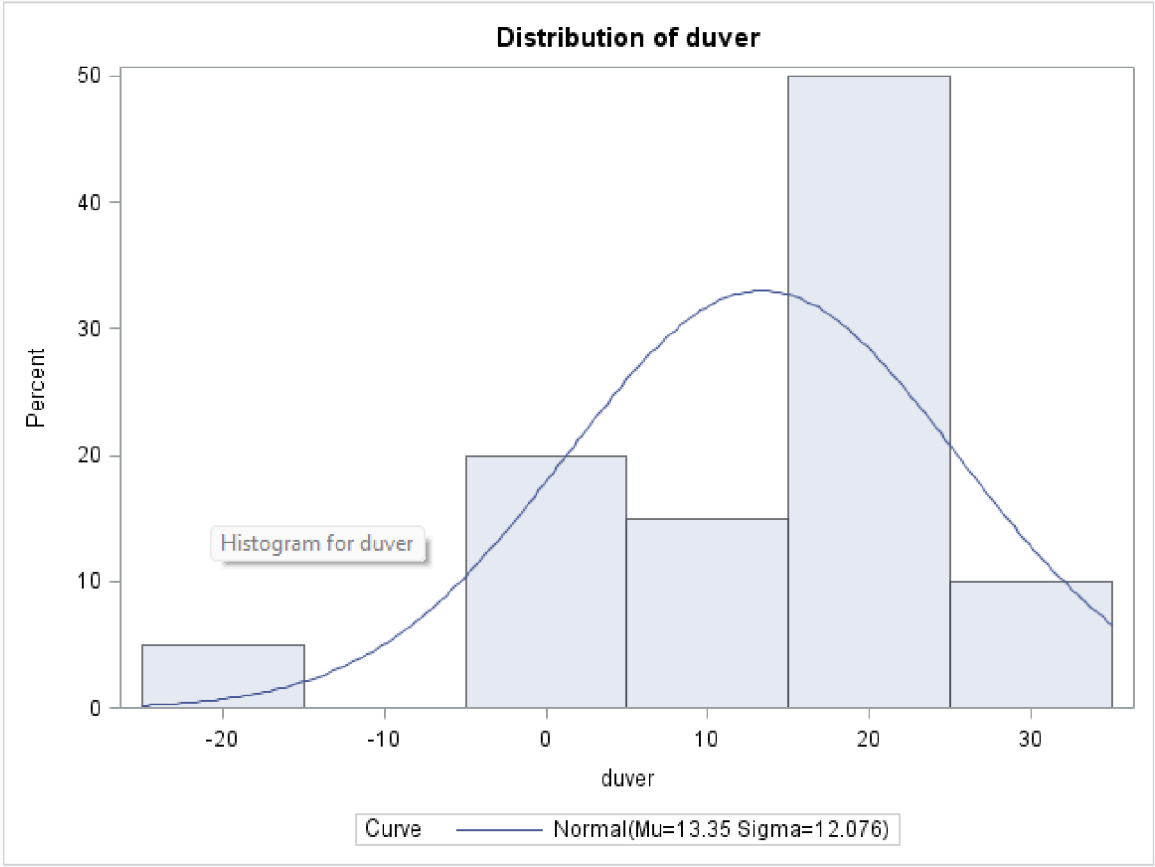
Kvantilový rozděl ukazuje distribuci dat. Mediánová hodnota (50 %) je 17.5, což znamená, že polovina pozorování má hodnotu "důvěr" nižší než toto číslo.

Krajní pozorování (1% a 100%) ukazují minimální a maximální hodnoty v datech, které jsou -16.0 a 30.0.

Quantiles (Definition 5)	
Level	Quantile
100% Max	30.0
99%	30.0
95%	29.5
90%	26.5
75% Q3	21.5
50% Median	17.5
25% Q1	5.0
10%	-2.5
5%	-10.0
1%	-16.0
0% Min	-16.0

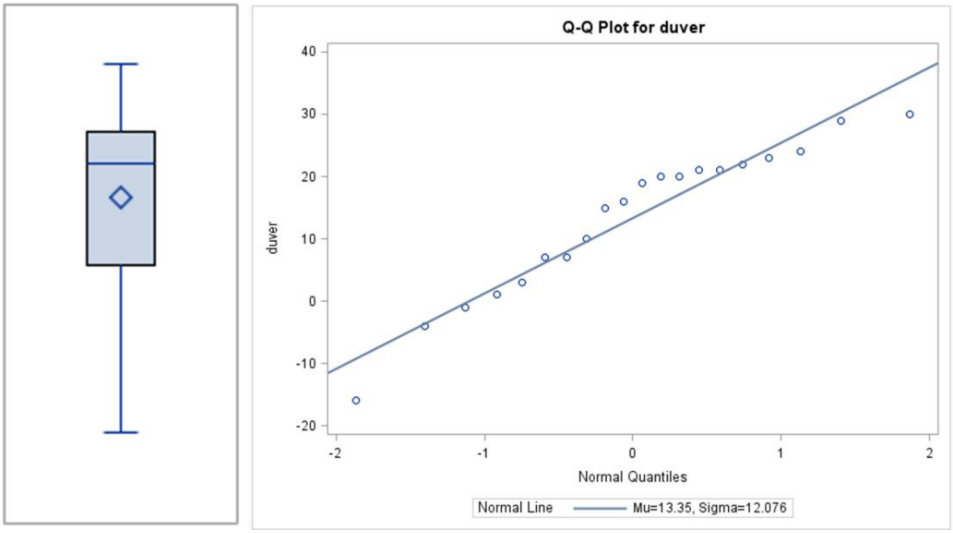
Obrázek 18: SAS. Kvantily pro duver. Zdroj: vlastní zpracování

Histogram znázorňuje rozdělení frekvencí hodnot proměnné „důvěr“. Graf mírně sleduje Gaussovi křivce, ale většina hodnot se nachází v pravé části grafu, což ukazuje levostrannou šikmost.



Obrázek 19: SAS. Histogram pro duver. Zdroj: vlastní zpracování

Z Boxplotu je vidět ze medián se nahází v horní části krabice, což potvrzuje levostrannou šikmost. Q-Q Plot ukazuje, že data sledují referenční přímku odpovídající normální distribuci, s určitými odchylkami, zejména v levém dolním rohu, což odpovídá levostranné šikmosti.



Obrázek 20: SAS. Boxplot a Q-Q Plot pro duver. Zdroj: vlastní zpracování

Na základě získaných dat neexistují důvody pro odmítnutí hypotézy o normalitě rozdělení dat pro proměnnou „důvěr“.

4.3.1.2 Test normality pro zabarvení

P-hodnota Shapiro-Wilkův testu je 0.0612 a je větší než 0.05, vzhledem k tomu neexistují dostatečné důvody pro zamítnutí normality rozdělení. Testy Kolmogorov-Smirnov, Cramer-von Mises a Anderson-Darling rovněž ukazují p-hodnoty vyšší než 0.05, potvrzujíce předpoklad o normalitě rozdělení dat.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.909061	Pr < W	0.0612
Kolmogorov-Smirnov	D	0.155693	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.073132	Pr > W-Sq	0.2444
Anderson-Darling	A-Sq	0.548574	Pr > A-Sq	0.1420

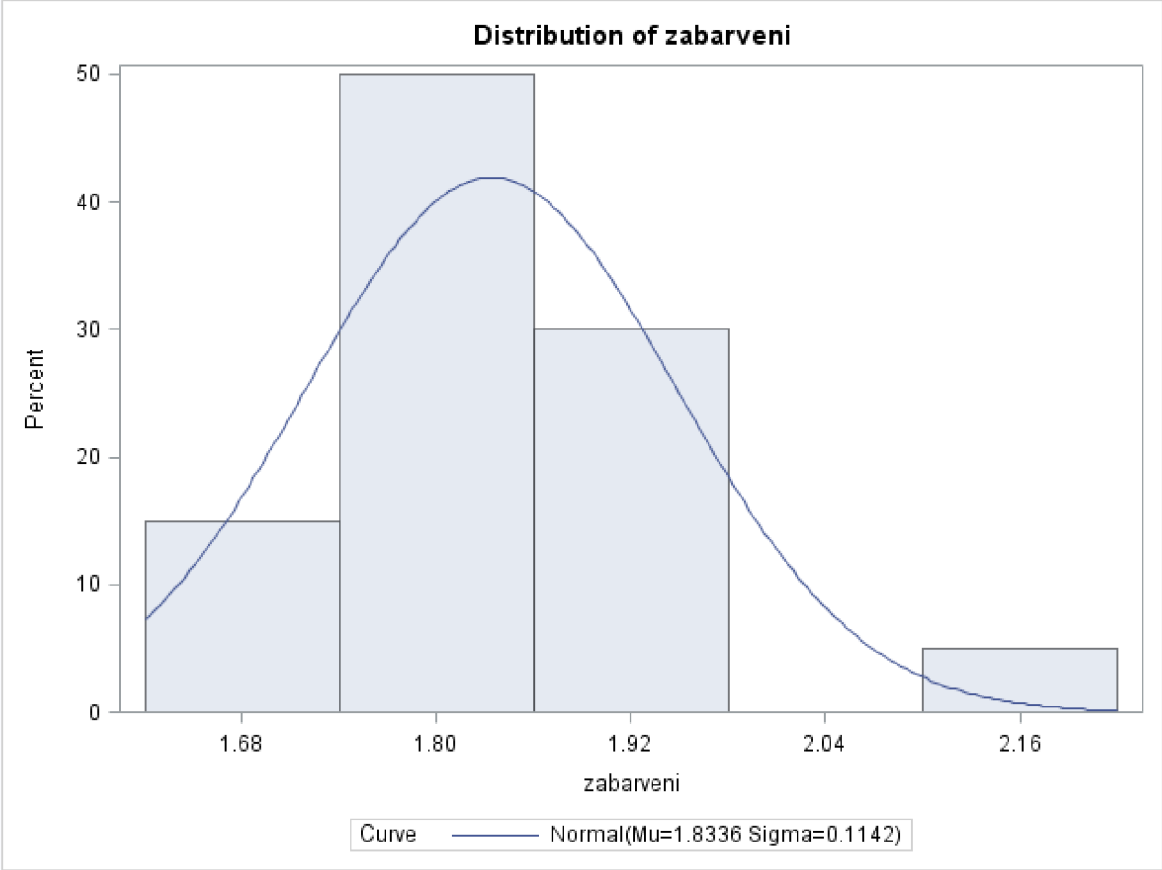
Obrázek 21: SAS. Test normality pro zabarveni. Zdroj: vlastní zpracování

Mediánová hodnota (50%) je 1.84206, což znamená, že polovina všech pozorování má hodnotu menší nebo rovnou tomuto číslu. Maximální a minimální hodnoty činí 2.16777 a 1.65001, což ukazuje rozsah dat.

Quantiles (Definition 5)	
Level	Quantile
100% Max	2.16777
99%	2.16777
95%	2.04761
90%	1.91986
75% Q3	1.89645
50% Median	1.84206
25% Q1	1.76931
10%	1.67296
5%	1.65752
1%	1.65001
0% Min	1.65001

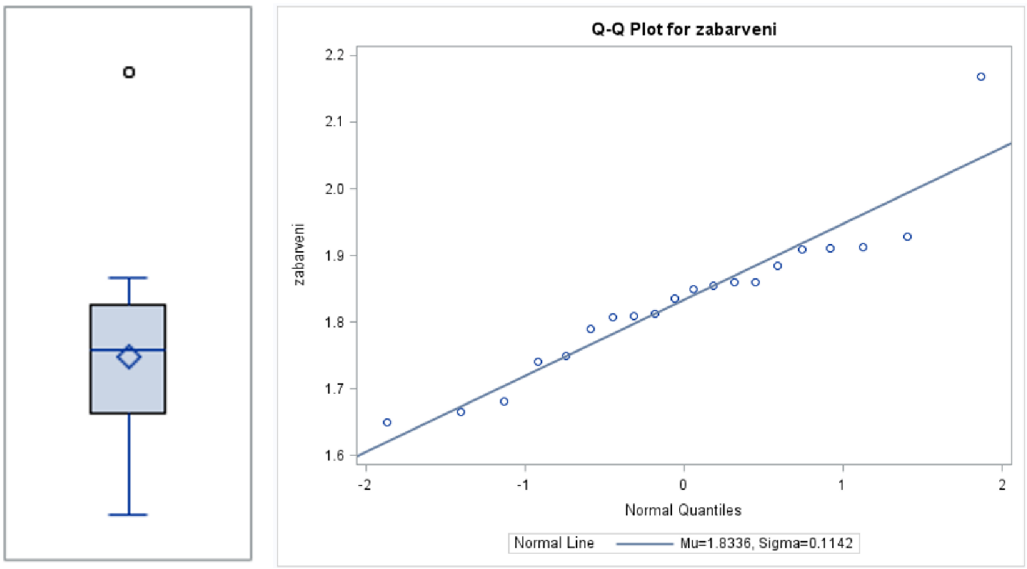
Obrázek 22: SAS. Kvantily pro zabarveni. Zdroj: vlastní zpracování

Histogram pro proměnnou "zabarvení" ukazuje že většina hodnot sleduje Gaussovi křivce, ale většina hodnot se nachází v levé části grafu, což ukazuje na mírnou pravostrannou šikmost.



Obrázek 23: SAS. Histogram pro zabarveni. Zdroj: vlastní zpracování

V boxplotu se medián nachází blíže k dolní části boxu, což potvrzuje mírnou pravostrannou šikmost. Boxplot také vykazuje jednu extrémní hodnotu. Q-Q Plot ukazuje, že data sledují referenční přímku odpovídající normální distribuci, s určitými odchylkami, zejména v pravém horním rohu, což odpovídá pravostranné šikmosti.



Obrázek 24: SAS. Boxplot a Q-Q Plot pro zabarveni. Zdroj: vlastní zpracování

Na základě získaných dat neexistují důvody pro odmítnutí hypotézy o normalitě rozdělení dat pro proměnnou „zabarveni“.

4.3.2 Regresní analýza

K provedení regresní analýzy používá příkaz REG. Parametr model specifikuje závislou proměnnou (důvěr) a nezávislou proměnnou (zabarveni).

```
proc reg data = mojetabulka;
model duver=zabarveni/r influence spec;
run;
```

Obrázek 25: SAS. Procedura pro provedeni regresní analýzy. Zdroj: vlastní zpracování

- /r — výpočet stundentizovaných reziduí a Cookovy vzdálenosti
- influence — míry vlivu každého pozorování na analýzu (výpočet charakteristik „leverage“ a DFFITS,)
- spec —ověřuje, zda je model adekvátně specifikován (Whiteov Test).

Výsledky této procedury zahrnou statistiky regrese, včetně koeficientů pro proměnnou zabarveni, hodnoty t-statistiky a p-hodnoty pro testování významnosti těchto koeficientů. Navíc budou poskytnuty míry adekvátnosti modelu, jako je R-Square, a grafy pro vizuální analýzu adekvátnosti modelu a ověření předpokladů regresní analýzy.

4.3.2.1 Analýza rozptylu (ANOVA)

Z tabulky analýzy rozptylu je vidět, že p-hodnota testu významnosti regresní funkce je 0.6795 a vyšší než hladina významnosti 0.05, což ukazuje na absenci statisticky významného vlivu nezávislé proměnné zabarveni na závislou proměnnou důvěr.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	26.88086	26.88086	0.18	0.6795
Error	18	2743.66914	152.42606		
Corrected Total	19	2770.55000			

Obrázek 26: SAS. Analýza rozptylu. Zdroj: vlastní zpracování

4.3.2.2 Těsnost závislosti

Těsnost závislostí určuje koeficient determinace R-Square, a rovná se 0.0097. z toho vyplývá, že model je vhodný pro popis závislosti na 0.97 %, což není statistické významné a potvrzuje absenci vlivu proměnné „zabarveni" na proměnnou „důvěr".

Root MSE	12.34610	R-Square	0.0097
Dependent Mean	13.35000	Adj R-Sq	-0.0453
Coeff Var	92.48011		

Obrázek 27: SAS. Těsnost závislosti. Zdroj: vlastní zpracování

4.3.2.3 Odhady parametrů lineárního regresního modelu

Tabulka ukazuje, že hodnota absolutního členu „Intercept" a parametr „zabarveni" nemají statisticky význam, protože oba parametry mají hodnoty testu významnosti vetší než 0.05 (0.9010 a 0.6795), což potvrzuje že model není statistické významný.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5.74444	45.55270	-0.13	0.9010
zabarveni	1	10.41349	24.79730	0.42	0.6795

Obrázek 28: SAS. Odhady parametrů lineárního regresního modelu. Zdroj: vlastní zpracování

4.3.2.4 Whiteov Test

Whiteov test ukazuje, že rezidua mají stejný rozptyl. P-hodnota je větší než 0, 05, což znamená, že model adekvátně specifikován.

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
2	2.55	0.2795

Obrázek 29: SAS. Whiteov test. Zdroj: vlastní zpracování

4.3.2.5 Vlivné, vybuchující a odlehlé pozorovaní

Vlivné pozorovaní je možné určit pomocí sloupce Cook's D nebo sloupce DFFITS tabulky Output Statistics. Kritérium pro posouzení celkové míry vlivnosti pozorování u sloupce Cook's D je p/n, což je 4/2 = 0.2. Hranici překračuje pozorování č.2, č.4 a č.19. Jedná se tedy o vlivné hodnoty. Tyto hodnoty budou odstraněny.

Hranice u sloupce DFFITS pro určení míry vlivnosti vyrovnané hodnoty se vypočítá jako absolutní hodnota $2\sqrt{p/n}$, což je $2\sqrt{2/20} = 0.63245$. Opět tuto hranici překračují pozorování č.2, č.4 a č.19 a také pozorovaní č.20.

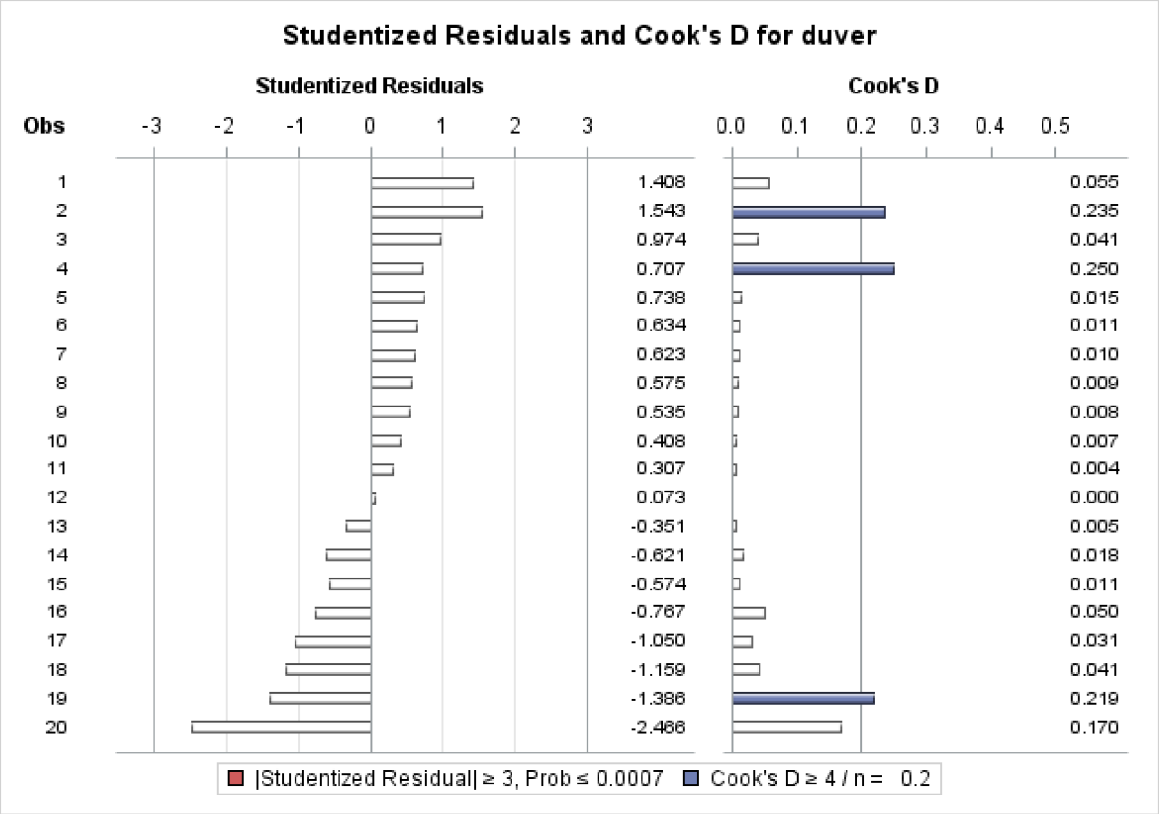
Odlehlá pozorovaní je možné stanovit pomoci sloupce Student Residual. Absolutní hodnota odlehlé hodnoty musí být větší než 2. Hranici překračuje pozorování č.20.

Vybočující hodnoty je možné určit pomoci sloupce Hat Diag H. Hranice pro vybočující hodnotu je 2*(p/n), což je 2*(2/20) a to se rovná 0.2. Tuto hodnotu překračuje pozorování č.4.

Pro dosazeni souboru bez vlivných pozorovaní tento postup byl opakován 4krat, a bylo odstraněno celkem 7 pozorovaní.

Output Statistics													
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS	
												Intercept	zabarveni
1	30	13.0759	2.8368	16.9241	12.016	1.408	0.055	1.4511	0.0528	0.9371	0.3426	0.0989	-0.0788
2	29	11.5943	5.0101	17.4057	11.284	1.543	0.235	1.6092	0.1647	1.0107	0.7145	0.6190	-0.5962
3	24	12.4646	3.4738	11.5354	11.847	0.974	0.041	0.9722	0.0792	1.0926	0.2851	0.1864	-0.1730
4	23	16.8297	8.7338	6.1703	8.726	0.707	0.250	0.6969	0.5004	2.1212	0.6975	-0.6472	0.6618
5	22	13.1294	2.8102	8.8706	12.022	0.738	0.015	0.7282	0.0518	1.1119	0.1702	0.0419	-0.0318
6	21	13.3675	2.7610	7.6325	12.033	0.634	0.011	0.6234	0.0500	1.1280	0.1430	0.0065	0.0022
7	21	13.5081	2.7862	7.4919	12.028	0.623	0.010	0.6120	0.0509	1.1309	0.1418	-0.0106	0.0192
8	20	13.0917	2.8284	6.9083	12.018	0.575	0.009	0.5638	0.0525	1.1402	0.1327	0.0367	-0.0289
9	20	13.5618	2.8064	6.4382	12.023	0.535	0.008	0.5246	0.0517	1.1448	0.1225	-0.0147	0.0220
10	19	14.1525	3.3575	4.8475	11.881	0.408	0.007	0.3984	0.0740	1.1884	0.1126	-0.0583	0.0641
11	16	12.3781	3.6024	3.6219	11.809	0.307	0.004	0.2988	0.0851	1.2127	0.0912	0.0627	-0.0586
12	15	14.1341	3.3328	0.8659	11.888	0.073	0.000	0.0708	0.0729	1.2085	0.0198	-0.0101	0.0111
13	10	14.1689	3.3800	-4.1689	11.874	-0.351	0.005	-0.3424	0.0749	1.1954	-0.0975	0.0513	-0.0562
14	7	14.3271	3.6104	-7.3271	11.806	-0.621	0.018	-0.6097	0.0855	1.1740	-0.1864	0.1113	-0.1201
15	7	13.8743	3.0298	-6.8743	11.969	-0.574	0.011	-0.5634	0.0602	1.1496	-0.1426	0.0508	-0.0588
16	3	11.7596	4.6866	-8.7596	11.422	-0.767	0.050	-0.7578	0.1441	1.2257	-0.3109	-0.2619	0.2513
17	1	13.6219	2.8356	-12.6219	12.016	-1.050	0.031	-1.0536	0.0527	1.0429	-0.2486	0.0420	-0.0568
18	-1	12.8959	2.9649	-13.8959	11.985	-1.159	0.041	-1.1714	0.0577	1.0186	-0.2898	-0.1218	0.1057
19	-4	11.4380	5.3246	-15.4380	11.139	-1.386	0.219	-1.4251	0.1860	1.0990	-0.6812	-0.6028	0.5825
20	-16	13.6267	2.8382	-29.6267	12.015	-2.466	0.170	-2.9446	0.0528	0.5191	-0.6956	0.1202	-0.1615

Obrázek 30: SAS. Output Statistics. Zdroj: vlastní zpracování



Obrázek 31: SAS. Studentizované rezíduá a Cook's D pro duver. Zdroj: vlastní zpracování

4.3.3 Regresní analýza po odstraněni vlivných pozorovaní

Oproti základnému souboru, po odstraněni vlivných pozorovaní model vykazuje statistickou významnost. Z tabulky analýzy rozptylu je vidět, že nová p-hodnota testu významnosti regresní funkce je 0.0188 a menší než hladina významnosti 0.05, což potvrzuje statistický význam vlivu nezávislé proměnné zabarveni na závislou proměnnou důvěr.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	328.19457	328.19457	7.58	0.0188
Error	11	476.57466	43.32497		
Corrected Total	12	804.76923			

Obrázek 32: SAS. Analýza rozptylu upraveného souboru. Zdroj: vlastní zpracování

Koeficient determinace R-Square je 0.4078. z toho vyplývá, že nový model je vhodný pro popis závislosti na 40,78 %.

Root MSE	6.58217	R-Square	0.4078
Dependent Mean	16.69231	Adj R-Sq	0.3540
Coeff Var	39.43236		

Obrázek 33: SAS. Těsnost závislosti upraveného souboru. Zdroj: vlastní zpracování

Tabulka „Parameter Estimates“ ukazuje, že hodnota absolutního členu „Intercept“ a parametr „zabarveni“ mají statisticky význam po upraveni základního souboru, protože oba parametry mají hodnoty testu významnosti menší než 0.05 (0.0121 a 0.0188), což potvrzuje že model je statistické významný. Díky tomu je možné sestavit výslednou regresní funkci: Důvěr = 200.69128–99.17732*zabarveni.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	200.69128	66.87760	3.00	0.0121
zabarveni	1	-99.17732	36.03427	-2.75	0.0188

Obrázek 34: SAS. Odhady parametrů lineárního regresního modelu upraveného souboru. Zdroj: vlastní zpracování

Stejně jako v základním souboru Whiteov test potvrzuje shodu rozptylů.

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
2	1.21	0.5466

Obrázek 35: SAS. Whiteov test upraveného souboru. Zdroj: vlastní zpracování

4.3.4 Korelační analýza

K provedení korelační analýzy používá příkaz CORR, který vypočítává koeficienty korelace mezi páry proměnných.

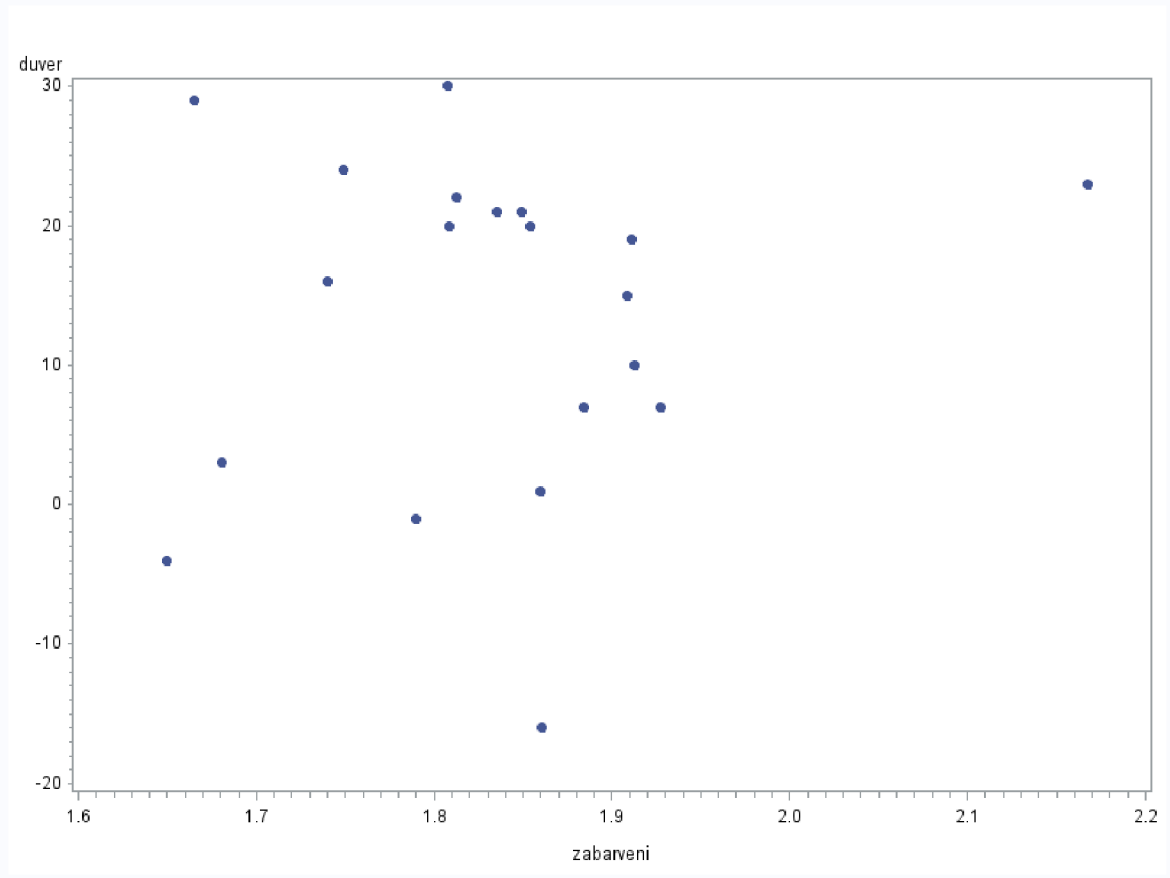
Parametry „pearson“ a „spearman“ provádí výpočet jak koeficientů korelace Pearsona, tak Spearmana. Koeficient korelace Pearsona měří lineární vztah mezi dvěma proměnnými, zatímco koeficient korelace Spearmana měří monotónní vztah (který může být i nelineární).

```
proc corr data = mojetabulka pearson spearman;
var duver zabarveni;
run;
```

Obrázek 36: SAS. Procedura pro provedení korelační analýzy. Zdroj: vlastní zpracování

4.3.4.1 Korelační analýza základního souboru

Z korelačního pole základního souboru nelze stanovit lineární závislost, ani rostoucí nebo klesajících tendence, což potvrzuje dříve stanovenou přítomnost odlehlých, vybočujících a extrémních hodnot a zároveň absenci statistické významnosti modelu.



Obrázek 37: SAS. Korelační pole. Zdroj: vlastní zpracování

Vzhledem k absenci lineární závislosti v základním souboru, pro stanoveni korelačního koeficientu byl použít Spearmanův korelační koeficient.

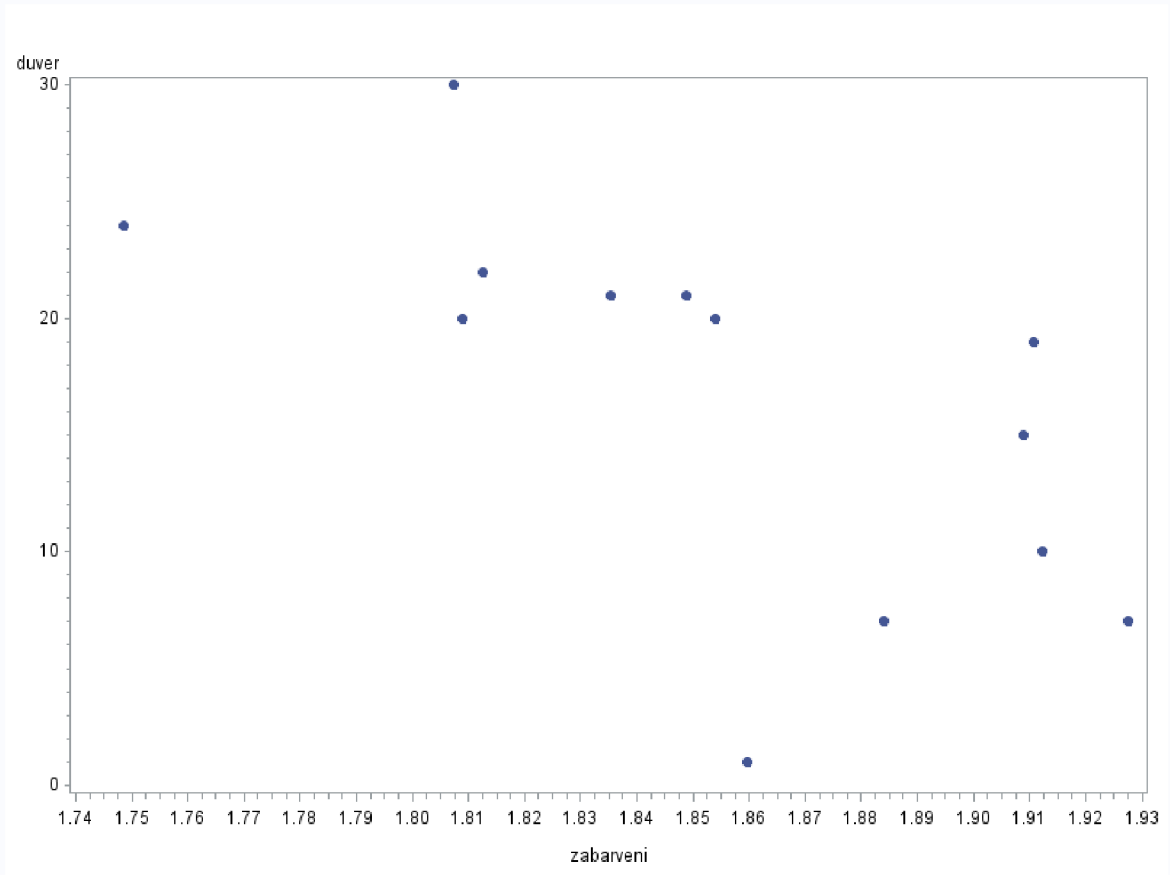
Spearman Correlation Coefficients, N = 20 Prob > |r| under H0: Rho=0		
	duver	zabarveni
duver	1.00000	-0.09259 0.6978
zabarveni	-0.09259 0.6978	1.00000

Obrázek 38: SAS. Spearmanův korelační koeficient. Zdroj: vlastní zpracování

- Spearmanův koeficient základního souboru je -0.09259, což je velmi slabá negativní závislost a není statistické významná.

4.3.4.2 Korelační analýza upraveného souboru

Narozdíl od základního souboru, z korelačního pole upraveného souboru lze stanovit slabou klesající tendenci a nižší stupeň lineární závislosti.



Obrázek 39: SAS. Korelační pole upraveného souboru. Zdroj: vlastní zpracování

Vzhledem k odstraněni extrémních hodnot, pro stanoveni kořalečního koeficientu upraveného souboru byly použity Spearmanův a zároveň Pearsonův korelační koeficienty.

Pearson Correlation Coefficients, N = 13 Prob > |r| under H0: Rho=0		
	duver	zabarveni
duver	1.00000	-0.63860 0.0188
zabarveni	-0.63860 0.0188	1.00000

Spearman Correlation Coefficients, N = 13 Prob > |r| under H0: Rho=0		
	duver	zabarveni
duver	1.00000	-0.81932 0.0006
zabarveni	-0.81932 0.0006	1.00000

Obrázek 40: SAS. Spearmanův a Pearsonův korelační koeficienty upraveného souboru. Zdroj: vlastní zpracování

- Pearsonův korelační koeficient je -0.63860, což je mírná, negativní závislost. Je statistické významná.

- Spearmanův korelační koeficient je -0.81932, což je silná negativní závislost. Je statistické významná.

4.3.5 Výsledky analýzy

Vzhledem k velkému počtu extrémních hodnot není možné říct, že odstraněné pozorovaní opravdu nenesou v sobě statistický význam. Proto není možné používat lineární vztah mezi proměnnými za prokázány, i když po upraveni souboru došlo k vzniku silné negativní závislosti mezi proměnnými.

Ale je možné říct, že pozitivnější zabarveni článku mírně negativně ovlivňuje úroveň důvěru, protože v obou případech Spearmanův korelační koeficient je negativní, i když velmi slabý.

5 Závěr

Hlavním cílem této bakalářské práce je posoudit o vliv emočního zabarvení článků na důvěru veřejností k zpravodajským portálům.

V teoretické části práce byly zkoumány základní principy, historie vývoje a metody provádění analýzy sentimentu na zaklade existujících výzkumů a odborné literatury. Popsány metody analýzy sentimentu, kde hlavním nástrojem je knihovna Stanford CoreNLP. Dále jsou popsány metody statistické analýzy pomoci programu SAS.

V praktické části práce byl prováděn sběr dat k analýze na zaklade existujícího studia. Následně byla vytvořena aplikace pomoci programovacího jazyka Java s využitím knihovny Stanford CoreNLP a prováděna analýza sentimentu získaných dat. Výsledky analýzy sentimentu byly probrány a porovnány s úrovním důvěry k zpravodajským portálům pro následnou statistickou analýzu. Byla prováděna statistická analýza pomoci programu SAS a dosažen závěr o vztahu mezi emočním zabarvením článků a úrovní důvěry vůči zpravodajským portálům, což odpovídá hlavnímu cíli práce.

Přestože výsledky práce naznačují určitou souvislost mezi mírou důvěry ve zpravodajské portály a emočním zabarvením článků, tato souvislost není přímá a nehraje klíčovou roli při stanoveni úrovně důvěry veřejnosti.

6 Seznam použitých zdrojů

BING Liu 2012. Sentiment Analysis and Opinion Mining. Springer Nature Switzerland AG 2012. ISBN 978-3-031-01017-0.

CAMBRIA, E., & HUSSAIN, A. (2015). Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis. Springer. ISBN 978-3-319-23653-7.

D'ANDREA, A., FERRI, F., GRIFONI, P., & GUZZO, T. (2015). Approaches, tools and applications for sentiment Analysis implementation. International Journal of Computer Applications, 125(3), 26–33. https://doi.org/10.5120/ijca2015905866.

STANFORD NLP GROUP. CoreNLP. [online] [cit.] Dostupné z: https://stanfordnlp.github.io/CoreNLP/index.html

ABDISHAKUR Hasaan. What Is Statistical Analysis [online] [cit. 21.12.2022] Dostupné z: https://builtin.com/data-science/statistical-analysis
SAS. About Sas. [online] [cit. 25.02.2024] Dostupné z: https://www.sas.com/cs_cz/company-information/why-sas.html

SAS. Company History. [online] [cit. 25.02.2024] Dostupné z: https://www.sas.com/cs_cz/company-information/history.html

SASNRD. Scatter Plot with PROC SGPLOT. [online] [cit. 25.02.2024] Dostupné z: https://sasnrd.com/sas-scatter-plot-example/

SASNRD. Histogram With PROC SGPLOT. [online] [cit. 25.02.2024] Dostupné z: https://sasnrd.com/sas-histogram-example/

SASNRD. Box Plot With PROC SGPLOT. [online] [cit. 25.02.2024] Dostupné z: https://sasnrd.com/sas-box-plot-example/

SAS INSTITUTE INC. Creating a Normal Probability Plot. [online] [cit. 04.08.2024] Dostupné z: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.3/qcug/qcug_capability_sect250.htm

CODY, R. P., SMITH, J. K. (2006). Applied Statistics and the SAS programming language. Pearson. ISBN 978-0131465329

DATATAB TEAM. Normality test. [online] [cit. 27.02.2024]. Dostupné z: https://datatab.net/tutorial/test-of-normality

ELLIOTT, A. C., WOODWARD, W. A. (2015). SAS Essentials: Mastering SAS for Data Analytics. John Wiley & Sons. ISBN 978-1119042167

SANDERS Linley. Trust in Media 2023: What news outlets do Americans trust most for information? [online] [cit. 08.05.2023]. Dostupné z:

https://today.yougov.com/politics/articles/45671-trust-in-media-2023-what-news-outlets-trust-poll

7 Seznam obrázků, tabulek, grafů a zkratek

7.1 Seznam obrázků

7.2 Seznam tabulek