UNIVERZITA PALACKÉHO V OLOMOUCI

Filozofická fakulta

Katedra anglistiky a amerikanistiky

Machine Analysis of Indicators of Tabloidization in Wall Street Journal (2000 – 2020)

Diplomová práce

Autor: Bc. Jan Ranostaj

Vedoucí práce: Mgr. Ondřej Molnár, Ph.D.

Olomouc 2021

UNIVERZITA PALACKÉHO V OLOMOUCI

Filozofická fakulta

Katedra anglistiky a amerikanistiky

Machine Analysis of Indicators of Tabloidization in Wall Street Journal (2000 – 2020)

Strojová analýza indikátorů bulvarizace v deníku Wall Street Journal (2000 – 2020)

Diplomová práce

Autor: Bc. Jan Ranostaj

Vedoucí práce: Mgr. Ondřej Molnár, Ph.D.

Olomouc 2021

Místopřísežně prohlašuji, že jsem diplomovou práci vypracoval samostatně pod odborným dohledem vedoucího diplomové práce a uvedl jsem všechny použité podklady a literaturu.

V Olomouci dne 13.12.2021 Podpis……..

Poděkování

Rád bych poděkoval vedoucímu diplomové práce Mgr. Ondřeji Molnárovi, Ph.D. za jeho neuvěřitelnou trpělivost, ochotu a cenné rady.

List of Abbreviations

WSJ – Wall Street Journal

Web Scraping Tool – Crawler

Processing Tool – Analyzer

TABLE OF CONTENT

1 INTRODUCTION

The issue of tabloidization became a heavily discussed topic in recent years. With the invention of internet and subsequent digitalization of news, news publishers had to adjust to the new paradigm and learn how to utilize various tools and approaches to survive and show profit in the everchanging landscape of our concurrent societal and economic environment (Sellers 2006). In this thesis, we will explore one of those tools – **tabloidization**.

Defined by Bird as "stylistic and content changes in journalism, usually perceived as representing a decline in traditional journalistic standards" (2009, 40), tabloidization itself is a complex term, and its comprehension necessitates consideration of various contexts, as it depends heavily on historic, cultural, and economical settings in which it is examined. Due to these factors, to try to analyze the process of tabloidization is rather difficult, and various researchers have produced studies with only partial or unconvincing evidence of tabloidization (Bek 2004, Uribe and Gunter 2004, Esser 1999), although there is a consensus that various British broadsheets are indeed undergoing a notable process of tabloidization to stay competitive (McLachlan and Golding 2000).

It is also apparent that up until this point most of the authors interested in the issue of tabloidization were mostly focused on a dataset of limited sample size, which might inevitable lead to some issues, as changes within newspaper that are often linked to the process of tabloidization are necessarily subjected to current events – e.g., change of the owner of the newspaper or its business strategy (to appeal to a larger or different audience), seasonal trends (elections), current national as well as global events (pandemic, economic crisis), etc. This necessarily leads to a certain amount of variance being introduced to the research, and while the individual authors usually try and take this into the consideration, it still represents a limiting factor.

In this thesis we decided to focus on the quantitative analysis of the indicators of tabloidization in case of one of the largest American broadsheets – **Wall Street Journal** (WSJ). The technological progress not only forced newspaper publishers to adjust their business strategies to survive as the circulation of both UK and US newspapers continuously declines (Ahrens 2016), but it also supplied us with new tools and possibilities to gather and process data. In this thesis, we would

like to utilize various programming tools and resources – mostly based on programming language **Python** – to gather the dataset of WSJ articles published in the 21-year-long period of 2000 – 2020, which consists of approximately one million articles, to detect the potential indicators of tabloidization and try to spot potential trends towards increasing tabloidization throughout the time. While qualitative studies are often criticized for their lack of objectivity and generalizability (Myers 2000, 9), the quantitative research will necessarily lack in terms of quality, as its findings "are not tested to discover whether they are statistically significant or due to chance" (Ochieng 2009, 17). Consequently, while we will do our best to optimize the data collection process as well as the computer analysis process to be as conclusive as possible, the goal of the thesis is not necessarily to provide a flawless tabloidization analysis tool, but rather try to showcase a potential trend towards tabloidization, and to answer a question: *"Is one of the largest broadsheets in the world showing a trend towards of tabloidization?"*

1.1 STRUCTURE OF THE THESIS

In the following chapters, we will briefly introduce both **broadsheets** and **tabloids** and subsequently set up a theoretical framework in which we will explore the issue of tabloidization, including its origin and development over time, as well as its dependency on a technological, and cultural context it finds itself in. The role of the internet on the development of current media will also be considered in this section. Since tabloidization is usually linked to a decrease of journalistic standards, and therefore carries certain negative connotations, we will also explore both the positive and negative impact of tabloidization on the newspaper and its readers.

In the next part of the thesis focused on the methodology, dimensions of the tabloidization will be defined in which potential indicators of tabloidization will be introduced. We will then define our own criteria based on which we will select several of these indicators for our computer analysis.

In the practical part, we will utilize data we gathered from the WSJ online archive[1] via web scraping[2] program (**crawler**). We did our best to collect the data set in a form that is as consistent as possible and ready for a computer analysis. This data set will be then used as an input for a program (**analyzer**) written in Python that will analyze chosen indicators of tabloidization. The output of the program will then be interpreted (including a graphical run chart).

In the final part of the thesis, we will try and interpret the reached results and compare them with findings of other authors.

[1] Source: https://www.wsj.com/news/archive/years
[2] Web scraping is data collection process utilizing a bot or a "crawler" to extract data from websites

2 ABOUT TABLOIDIZATION

In the following section, we will briefly introduce broadsheets and tabloids and the development of the process of tabloidization. Subsequently, we will explore the factors that cause tabloidization. Finally, we will shortly discuss the impact of tabloidization – both negative and positive. This will allow us to properly introduce the concept so we can set up our methodological framework in the following chapter.

2.1 BROADSHEETS VS TABLOIDS VS INFOTAINMENT

In our everyday live we generally distinguish two types of newspapers, **broadsheets** (quality press) and **tabloids** (yellow press). Even though the nomenclature relates to their respective sizes, this dichotomy carries certain connotative meaning regarding not only the form of their layout, but also their reputability, the content they chose to focus on, target audience, and style of writing.

2.1.1 Broadsheets

Collins Dictionary defines *broadsheet* as "a newspaper that is printed on large sheets of paper, [that is] generally considered to be more serious than other newspapers"[3]. Hughes marks as the first broadsheet the journal *Courante uyt Italien, Duytslandt, &c* that emerged in Netherlands in 1618, since before that news were being published sporadically as a reaction to a singular event (2021). Their large format was attributed to the tax imposed on number of pages in Britain. Typical British Broadsheet measures a 29 ½ inches (74 cm) in length when unfolded, and 23 ½ inches (60 cm) wide. The traditional American broadsheet is slightly longer and narrower (30 x 22 ¾ inches or 76 x 58 cm). The dimensions will usually refer to the size of the newspaper when folded on a newsstand.

Broadsheets have always been considered more reputable of the two, targeting "major" topics or news related to economics, politics, and significant

[3] Source: https://www.collinsdictionary.com/dictionary/english/broadsheet

national events, dedicating majority of the space to the text and informative content, written in an objective, informative style. They tend to include foreign news and affairs and generally appeal to more educated, economically, and politically interested segment of the reader market.

In US almost all major newspaper remaining are broadsheets, such as New *York Times* and *Washington Post*, or the most widely circulated paper in the country *USA Today* (Hughes 2021). In Britain the process of tabloidization was much more significant, and most of the major circulating newspapers are tabloids. *Daily Telegraph* and *Financial Times* are some of the examples of broadsheet that are still present in the UK to this day. Notably, a lot of traditional broadsheets both in US and UK have downsized their format over the years (WSJ included).

2.1.2 Tabloids

According to Fang, the term *tabloid* itself – based on easy to swallow dose of medicine – first appeared in London in the early 20th century and referred to news printed in a format that could be comfortably held in one hand and easily read in an automobile (1997, 103). Compressed newspapers of such a type used to focus more on entertainment, and the term referred not only to the format of the newspaper, but its content as well (ibid., 104). Hughes marks the year 1901 as an origin of tabloids, when Alfred Harmsworth (founder of the *Daily Mail* in England) was invited to edit *New York World*[4] to test his theory that smaller formats would be more efficient for both writers and readers (2021). In 1903 *The Daily Mirror* in the UK was launched as the first tabloid and *The Sun* and *The Daily Star* eventually followed.

In the UK we distinguish two types of tabloids. **Red-tops** like *The Sun* are characterized by large masthead[5] and report on politics and international news but tend to include more celebrity gossip and scandals, while the **middle-market dailies** like the *Daily Mail* are somewhere between the broadsheets and red-tops. In the US and Canada, the published tabloids are so-called *supermarket tabloids*, as

[4] Newspaper published in New York City from 1860 until 1931
[5] Large font title at the top of a newspaper front page containing the newspaper's title. (source: www.bbc.co.uk/bitesize/guides/zps4qty/revision/1)

they are usually sold in the cash-out line. These tabloids utilize large, catchy titles and other aggressive strategies to capture the attention of the buyers.[6]

Even though tabloids can be beneficial, as they supply the reader with accessible, easy-to-digest type information, and as such offer content more relaxing than serious broadsheets, the term usually carries a negative connotation. According to Rowe "There are few words (both adjectives and nouns) in the field of news and Journalism Studies that are likely to attract more viscerally negative responses than 'tabloid', and few processes more commonly used to signify the decline of the contemporary news media than tabloidization" (2009, 350).

In UK most of the large circulation newspaper are tabloids, such as *The Sun*, *Daily Star* or *Daily Mirror*, while the situation in the US is opposite, with most of the large newspapers being broadsheets. *The Globe* or *Daily Enquirer* are some of the examples of US tabloids.

2.1.3 Infotainment

Naturally, newspapers are rarely classified as either pure tabloid or a pure broadsheet, as the dichotomy represents a spectrum rather than a clear-cut division. Consequently, most of the newspapers find themselves somewhere on this spectrum. This gave a raise to a modern term – **infotainment.** Infotainment (a combination of words *information* and *entertainment*) represents a way in which informative news are presented in an entertaining manner, often utilizing images or audio to accompany articles, sensationalism, satire, and other attention-capturing devices. Thusu talks about "a tension between informing and educating the public and entertaining the crowd in the marketplace" (2008, 15). While originally referring to mostly television, the term became applicable to online news as well. Infotainment is a term that should be mentioned while exploring tabloidization, although the term itself should be in this context considered to be more of an indicator of tabloidization.

[6] Source: www.bbc.co.uk/bitesize/guides/zps4qty/revision/1

2.2 DEVELOPMENT OF TABLOIDIZATION

"Long before the term was actually coined, tabloidization was a focus of criticism and concern, that began with the emergence of more popular journalistic formats, such as 'penny-press' of the 1830s, in which writers produced dramatic, human interest news of crime and mayhem, frequently with implied or overt moral" (Bird, 41). Bird mentions that the term *tabloidization* itself became popular quite recently (in the 1980s) but the process itself – representing decline in traditional journalistic standards – has been discussed for at least a century (2009, 40). Other authors define tabloidization slightly differently. In comparison, Conell understands the term as a series of processes that transform nationalist discourses into sensationalist discourses (1998, 12). For the sake of this thesis, however, we will understand tabloidization as a process of *shifting traditionally quality press standards towards yellow press standards*.

Even in its beginnings, the shift towards tabloidization seems to be linked to a technological advancement, as according to Fang, the initial change of printed format could be attributed to newspaper publishers trying to adjust to a technological shift from horse-drawn buses to trolleys and subways as there was a need for a piece of newspaper that could be comfortably read with one hand while standing and holding a strap-hanger with the other one (1997, 103).

In the US, the beginnings of tabloidization can be traced to Yellow Journalism in 1890s (Esser 1999, 293), although the term itself did not become popular until the launch of the newspaper *USA Today* in 1982 (S. E. Bird 2009, 42). The concept of tabloidization is slightly different in the US as well, as US does not have such a clear dichotomy between broadsheets and tabloids as the UK. Esser states that one of the reasons is the invention of radio, which caused advertisers to move from written media as their first choice to radio as a means of reaching mass audience (1999, 295).

Part of the reason why the exact definition of tabloidization is not clearly set is because the term might take on a different meaning in a different cultural setting. According to Bird, people in different parts of the world might understand tabloidization differently, mostly due to the different societal values and norms (2009, 92). While the British and American media have the longest histories of tabloids (Esser 1999, 294), the process took on a different form in US then it did in

Great Britain. Bird also claims that it is necessary to consider cultural specifics when discussing tabloidization (42). For example, since Americans are less open about their sexuality, and favor gore, crime is the number one topic in the US (Esser 1999, 295).

In Great Britain, tabloidization first occurred when *establishing Daily Mirror*. Britain has a long history of tabloids, and tabloids were historically more successful here than in the US. The dichotomy between tabloids and broadsheets is also much clearer than in the US, with *The Sun*, *Daily Mirror* and *Daily Mail* being some of the textbook examples of the tabloid. Esser points out that this prevalence of the tabloids when compared with the US is again caused by the radio, since in the UK radio was established as a social service and was not made available as an advertising platform. The advertisers therefore had to stick to the tabloids to reach the masses (1999, 295). Since the British are less religious and puritan then Americans, sex is the predominant topic in British tabloids compared to US (ibid., 313).

In case of UK, we can also find more evidence that the trend of tabloidization is indeed taking place. Golding and McLachlan found out that the amount of political news stories and their average length have become more similar between quality and tabloid newspapers (2000). Research conducted by Uribe and Gunter shows that the coverage in the UK is taking place especially when it comes to *form* and *style* – marked by increasing emphasis on the headlines, visuals, and personalization – but not as much in the content (2004).

In the US, there has not been as much research focused on tabloidization as in case of the UK, mostly because of the different nature of tabloids and tabloidization. For example, in her paper *News We Can Use: An Audience Perspective on the Tabloidisation of News in the United States*, Bird focuses on tabloidization in television rather than newspapers (E. Bird 1998).

As for non-Anglo-American countries, Esser states that for example in Germany the societal norms are vastly different, as Germans would never print a story about a politician's personal affairs, as Germans are much more confidential (1999). First, it's simply not part of their culture, and second, in Germany one's reputation and privacy are protected constitutionally, so tabloids will take on a slightly different form there than in US and UK. Similar conclusions were reached by Brandelid and Eklund when examining tabloidization in Scandinavia (2021).

2.3 CAUSES OF TABLOIDIZATION

2.3.1 Economic factors

Economic factors are some of the main forces driving tabloidization, especially in the current environment, as "print news agencies have been under pressure from falling sales and advertising revenue and increased competition" (Spillane a al. 2020, 1). Companies usually try to cope with **market competition** by increasing sales by utilizing so called four P's (product, place, promotion and price). Tabloidization might help with the process as the change of format might offer cheaper cost of both journalism and print cost as well as more effective marketing strategies. These changes often require change of the target audience – either by switching to tabloid format to change or broaden the segment of readers the newspaper is targeting or integrating a subscription fee to narrow the target segment. Traditionally quality press media might therefore resort to tabloidization in case they are desperately searching for profit. In such cases, it might make sense to make a risky shift towards tabloid or utilize larger amount of infotainment to boost sales to avoid inevitable bankruptcy. Spillane, however, calls this strategy a losing one in the long term (ibid.).

2.3.2 Technological factors (invention of internet)

While we already discussed some of the technological factors triggering tabloidization in the chapter dedicated to the development of the tabloidization, such as switch to trolleys and automobiles, requiring a smaller format of a newspaper, without a doubt the most impactful piece of technology during in the recent history was the invention of internet. The internet brought significant lowering of the barriers of entry for the potential competition, as producing news in the digital form is significantly less logistically and financially demanding than in case of its printed counterpart. This meant new opportunities for countless potential news producers trying to find their niche in the market. Since this meant that news readers had numerous new sources to pick from, according to Bird it

resulted in news producers mostly abandoning their attempts to charge for their content, as audiences would not pay for content they can get elsewhere (2009, 47)

This has caused newspaper producers to adjust their marketing strategy. Suddenly adds became the main source of revenue. The focus of online news producers turned towards making the reader visit the webpage and stay on the webpage rather than keep paying a subscription. This resulted in more emphasis on provocative headers and catchy thumbnails. While a witty and interesting header was always a part of advertising, in the digital age this practice was pushed to such an extreme that it led to the emergence of a new term *clickbait*, standing for a "a certain kind of web content advertisement that is designed to entice its readers into clicking an accompanying link" (Potthast, et al. 2016, 810) – a practice, which is generally frowned upon, but to some extent remain an important part of the digital journalism. The emphasis on timeliness, novelty, and shock value of the content seems to be more important than ever, as the plethora of sources meant that even the higher quality journal (spending more resources on journalism) would have to bring interesting and provocative news as fast as possible, as taking the time for the "deep journalism" would mean that the news reader would simply read it elsewhere for free.

Nowadays, this dynamic seems to be more present than ever, as a simple swipe to the left on a smartphone will lead to list of news presented to you based on an algorithm evaluating readers preferences, interests as well as popularity of individual sources. And while there are traditional broadsheets and high-quality newspapers that are still using the subscription-based model (WSJ and Guardian among others), they still compete in the very same environment.

2.3.3 Other factors

Tabloidization is of course caused by other factors besides *economic circumstances* and *technological advancements*. *Political changes* are often a factor that can affect tabloidization. Media being prohibited from reporting on personal lives of politicians in Germany can serve as one of the examples where potential change in a legislature could trigger tabloidization. On a greater scale, Jelínková points out Velvet Revolution in 1989 in Czech Republic as an example of political change

triggering tabloidization, as this change led to privatization of media and emergence of new media producers (2019).

Another common factor potentially triggering tabloidization is a change in company policy / core values. This might be a result of a change of an ownership or changes in company core values motivated by a hope for a larger profit in the future. WSJ went through such a change in 2016, when the company switched to a digital-first approach to adapt to the quickly changing environment of the modern world.

2.4 IMPACT OF TABLOIDIZATION

2.4.1 Negative impact of tabloidization

As we have stated, we consider tabloidization to be a decrease in professional standards of journalism, and its exactly lack of these standards that gives the tabloid its negative reputation. According to Jelínková, tabloids are often linked to low content quality and unethical methods of collecting information, and not adhering to the rules of quality journalism (2019, 16). Rowe states that tabloids depend heavily on the exploitation and amplification of fear (2009, 352). Esser perceives tabloidization as a process of downgrading hard news and upgrading sex, scandals, and infotainment (1999, 292).

One of the issues with tabloidization lies also in the fact that media have historically had a role of providing general population with relevant information based on which people can make well-informed political and economic decisions. This process might easily be affected by the dropping standards of news production. This seems nowadays more relevant than ever, as the average reader is consuming large amount of news from unregulated sources. Furthermore, disinformation seems to be on the rise, as the current society has to deal with previously unseen challenges, ranging from the US Election being affected by twitter (Smialer 2018), which might have changed the eventual outcome of the election, to Elon Musk claiming his fake news generating OpenAI company tech is too ‘scary’ to release (Baig 2019).

2.4.2 Positive impact of tabloidization

On the other hand, tabloids have historically been cheaper, more accessible, and convenient for the readers. They provide easy-to-get, easy-to-consume form of entertainment that does not require high level of education and offer consumers a form of relaxation. Tabloids often cover topics and issues that would not be covered in a quality press, and have no problem finding their target audience.

But what about the process of tabloidization itself? Bird states that it is important to consider impact of tabloidization in a context, since a movement more accessible news that speaks more directly to readers does not necessarily mean decline in standards (2009, 42) and that evidence suggests that a certain tabloidization of style – such as emphasis on storytelling – can actually help engage senses and emotions, and therefore enhance good journalism and claims that use of visual material makes the newspapers more appealing and allows newspapers compete with other form of media, such as television (ibid.). There is arguably nothing wrong with quality press borrowing couple pages from tabloids book to become more accessible to the masses, or to boost readability, such as utilizing the modern technology to provide readers with more audiovisual material to enhance storytelling.

3 METHODOLOGY

The goal of the following chapter is to set the framework in which we will examine tabloidization. For that we will need explore the dimensions of tabloidization that will serve as a theoretical base in which we can enumerate individual indicators utilized by various authors. We will then define our own criteria that will help us to cherry-pick several indicators that we will then programmatically process in our research. We will also briefly introduce the journal we will explore (WSJ) and the way we will conduct our analysis. Research of a similar nature produced by other authors will be also mentioned in this section. Finally, we will introduce the technological tools and frameworks which made our analysis possible.

3.1 DIMENSIONS OF TABLOIDIZATION

MacLachlan and Golding (2000) set the theoretical framework of tabloidization by dividing the process into three dimensions in which tabloidization could be explored and analyzed:

Range – or variety in terms of content – which examines how much space is given to the informative news (hard news) compared to space allocated to the entertainment (soft news). Ratio of political news / non-political news or domestic news / international news falls into this dimension.

Form – can be understood as simplifying the format of the newspaper. Number of images, size of titles, and emphasis on graphics fall into this dimension.

Style – style of writing. Increasing personalization of coverage belongs to this dimension.

MacLachlan and Golding's dimensions form a traditionally very decent and acknowledged basic framework in which various authors explored the issue of tabloidization (Esser 1999, Jelínková 2019). For the purposes of this thesis, we therefore chose to adhere to the three dimensions of **range**, **form** and **style** as defined by MacLachlan and Golding.

3.2 INDICATORS OF TABLOIDIZATION

In the following paragraph, we will enumerate various indicators presented by different authors in the context of the tree dimensions. This will establish a pool of indicators out of which we will choose the indicators for our analysis.

3.2.1 Range-related indicators

Increase of soft news / decrease of hard news

Wilzing and Seletzky define the **hard news** as news of high newsworthiness – usually related to politics or economics – that demand fast publication, while **soft news** do not require immediate publication and have low newsworthiness (human stories, gossip, etc.) (2010, 37). This dichotomy of soft news and hard news forms the most basic division when it comes to analyzing news. According to Bennett, part of tabloidization process is that "topics that were once relegated to gossip columns and the screaming headlines of the tabloids are now increasingly the stuff of mainstream news" (2016, 10). This also corresponds with the Esser's definition of tabloidization as a "downgrading of hard news and upgrading of sex, scandal, and infotainment" (1999, 292).

Increase of soft news (and decrease of hard news) is therefore one of the most prominent indicators of tabloidization. Esser calls this a change in newspaper perception of what they think voters need to know a part of decrease of journalistic standards (ibid. 293) and worsening of journalistic behavior (ibid. 299) and this indicator has been frequently explored by many authors interested in the research of tabloidization, such as Karlsson (2016), Steiner (2020) and Uribe and Gunter (2004).

Decrease of political and economic news

One of the main ways the decrease of hard news is manifested in is in decrease of political and economic news in favor of human stories, sport news, scandals, etc. Furthermore, the nature of the political and economic news might change as well, as the stories often become more personalized, or change the focus on the politicians' scandals instead of the newsworthy information. In his study about

tabloidization in Sweden Karlsson calls it a "shift from political to more lifestyle journalism" (2016, 151). Besides Karlsson, this indicator was explored by e.g., Esser (1999) or Rowe (2000)

Decrease of international news

Another manifestation of decrease of hard news takes form of increase of domestic news at the cost of international news. This fits nicely into the personalization aspect of the tabloid as the tabloid sacrifices potentially newsworthy international news at the cost of news that in a closer proximity to the readers. Together with the decrease of political news, the decrease of international news form the most prominent manifestation of this phenomenon. This indicator has been explored by Uribe and Gunter (2004) and Zapletalík (2020).

3.2.2 Form-related indicators

Decrease of the amount of text

Rowe claims that newspapers must operate with so-called space budget (2009, 353). Decrease in the amount of text is one of the primary form-related indicators when it comes to tabloidization, as this phenomenon simply must take place to make more room for larger, more provocative, colorful titles or more image material. Consequently, the reader has less available textual material to form an objective opinion. Jelínková (2019) and McLachlan and Golding (2000) are some of the authors that examined this indicator in their research.

Increase in the number of images

This decrease of the length of text is frequently the cost of increasing number of images. Tabloid images are often graphic, provocative, and personalized. The increase of the image material helps with the visual appeal of the newspaper, which comes at the cost of lower amount of information that gets to the reader. This indicator has been explored for example by Rowe (2009) and Zapletalík (2020).

Increase of audiovisuals to text ratio

While increase of the number of images comes almost always with the decrease of the amount of text in the written format of the newspaper, the same might not necessarily be true for the digital format where the space budget is technically unlimited. Due to this we will focus more on the **increase of audiovisuals to text ratio**, which comes at the cost of decrease of surface dedicated to the text**.** This forms a rather universally acknowledged indicator when analyzing tabloidization, as many other authors – such as MacLachlan and Golding (2000) use this indicator as well. While utilizing the space available for the article to add images that might provide additional information value might not necessarily be an indicator of tabloidization – especially in the digital context, as digitally published news do not suffer from the same space limitations as their printed counterpart – we will consider the increase of the space allocated to the images *at the cost of decrease of the space allocated to text* to be a reliable evidence of tabloidization even in digital world.

Increase in emphasis on graphics

Besides the increase of image material, increased emphasis on graphics – namely large and provocative subtitles full of exclamation marks and question marks and colorful elements present, especially on the front page represent a recognized indicator of tabloidization as well. The emphasis on graphics was thoroughly explored by Jelínková (2019).

3.2.3 Style-related indicators

Increase in personalization

Personalization is one of the most prominent indicators of tabloidization within the style dimension, as the goal of the tabloid is to make the story as relatable to the common user as possible. Personalization might be manifested through the focus of the article on private life of a prominent figure, such as politician or an actor, or telling the story from a perspective of a common person to make the story more relatable to the general audience. Esser was focusing on personalization when exploring the tabloidization within the German news media (1999).

Increase of vox-populi sources

One of the manifestations of personalization is decrease of professional sources, such as scientists, politicians, and economists in favor of quoting representants of general population. This increase in *vox-populi* sources is often quoted indicator of tabloidization. This indicator was examined by Bek (2004) and Brandelid and Eklund (2021).

Increase of negativity

For our research, we will also state increase of **negativity**. We have already stated that exploitation and amplification of fear (Rowe 2009, 352) is quite common tactic of tabloids. So is a trend of pessimism, pessimistic speculation, and scandals (Esser 1999). Heitzmann found out that negativity in the newspaper is positively linked to the user engagement, and therefore increases number of visits of the website (2020), which makes it a very relevant indicator in the digital environment, where number of visits and time spent on the website is the key due to the revenue from advertising.

Increase in dramatic sentiment of the title

Brandelid explored whether the sentiment of the title matches the sentiment of the article as an indicator of the tabloidization (2021). Jelínkova explored both **formal** aspects of the title (including colorfulness and size of the title), and **stylistic** aspects of the title (number of exclamation marks and question marks). This becomes especially relevant when the title does not match the content (viz the notion of the *clickbait* in the chapter 2.3.2). Heitzmann proved that increased negativity leads to increased user engagement on the website. Based on this we can determine that significant gap between the (negative) sentiment of the title and the (less negative) sentiment of the article content might have a luring effect and signal trend towards tabloidization.

Unequal representation in a conflict

Unequal representation of the two arguing parties when reporting a conflict is generally a sign of a tabloid, as it might serve the purpose of not letting one side defend itself, ergo creating a conflict where there is not necessarily one present. It is also generally a sign of biased journalism and step away from journalistic values. As such, it can be treated as an indicator of tabloidization. This indicator was examined by Jelínkova (2019) and Esser (1999).

Increase in use of expressive language

Use of expressive language is another feature typical for tabloid as it violates standards of journalism. This indicator has been analysed by Košnárová (2018) and Steiner (2020).

3.2.4 All indicators – summary table

RANGE	FORM	STYLE
Soft news increase / hard news decrease	Text length decrease	Personalization increase
Political news frequency decrease	Image count increase	Negativity increase
International news frequency decrease	Increase of surface dedicated to audiovisuals	Increase in dramatic sentiment of the title
	Increase in emphasis on graphics (titles)	Increase in vox populi sources
		Unequal representation in a conflict

Table 1: All indicators – summary

3.2.5 Indicators suitable for computer analysis

Now that we defined our indicators of tabloidization, our goal in this chapter will be to cherry-pick several indicators that are suitable for computer analysis. Not all these indicators are easily or accurately applicable to a machine analysis of such a large dataset. Therefore, our main factors of consideration when choosing our indicators for the analysis will be:

- whether or not its computer analysis is **feasible** – e.g., *equal representation of both parties* in a conflict is an indicator that cannot be easily analyzed programmatically.
- whether or not its computer analysis is **reliable** and sufficiently accurate – e.g., *expressive language* could be technically analyzed, but its accuracy would be skewed by the set of words we compare the dataset to, the style of writing varying from author to author or shifting with time, etc.
- whether or not they are **relevant** – e.g., *graphic emphasis in the title* is irrelevant in a digital archive, as most of the titles follow the same format.

Based on these three criteria alone, we can eliminate indicators that are not useful for our analysis. Under **feasibility**, we can safely eliminate indicators such as *frequency of international news*, *unequal representation in a conflict* and *frequency of vox populi sources*.

We can further eliminate certain indicators that could be processed programmatically, but their **accuracy** would be questionable at best, such as *expressive language*, *degree of personalization, amount of vox populi sources, etc.*

Finally, under **relevancy** criteria we can eliminate indicators such *emphasis on the graphics of the title* because we consider the indicator irrelevant in case of our dataset.

Based on the stated criteria of accuracy, relevancy, and reliability, we have selected following indicators for our analysis:

- **Text length increase (1)** – this is a good indicator of tabloidization, that can be easily analyzed. Furthermore, WSJ states the length of the article in words within the meta information present in each article. This information

is therefore readily available and very accurate. Even if not present, it is very easy programmatically count the number of words within the text.

- **Increase in amount of audiovisuals** – it is possible to process numerous audiovisual properties programmatically. WSJ again states the **number of images (2)** within the meta information, so we can accurately process this property. But it is also possible to programmatically process **number of videos increase (3)** within the article, or **image to text ratio increase (4)**.
- **Increase in negative sentiment of the article (5)** – While it is very difficult to programmatically process expressive language, there are publicly available tools utilizing AI and machine learning[7] to determine the degree to which the text is negative, positive, or neutral. We can apply these tools to determine whether there is a trend of increasingly negative sentiment within the WSJ articles over the years.
- **Increase in the negative sentiment of the title (compared to the article text) (6)** – While we cannot accurately measure dramatic aspect of the header, we can measure its sentiment using aforementioned tools and compare it to the sentiment of the text. A trend of increasing negative sentiment of the title varying significantly from the sentiment of the article could be considered an indicator of tabloidization.
- **Increase in number of articles containing symbols "!" or "?" in the title (7)** – Part of the dramatic sentiment of the title is established by the presence of question mark or quotation mark. We are storing the titles with our crawlers and we can very easily detect this parameter programmatically.
- **Decrease in number political or economic news (8) –** WSJ articles fall into certain sections based on the division of the main page. We can easily detect and store the section with our crawler and subsequently determine the percentage of news falling into these sections over the years.

[7] The two of the most popular sentiment analysis tools are **VADER** and **TextBlob**.

3.2.6 Indicator relevancy

When looking at these parameters, it's easy to spot that some of the potential trends are simply more relevant indicators of tabloidization than others. E.g., growing number of pictures is arguably less relevant indicator of tabloidization than increasing image to text ratio, as the latter might indeed mean the decline of text material compared to the audiovisual material, the former might mean that the journal is simply utilizing more audiovisual material, which might be caused by technological progress, and positively enhance consumer experience. We have decided to keep these less relevant indicators on purpose, as their analysis is relatively easy, and while they might not be strong indicators of tabloidization, their presence arguably signals that the WSJ is indeed utilizing more audiovisual material over time to keep up with the competition, which is arguably a good thing for a consumer. Lesser relevancy of some of these indicators will be mentioned and accounted for when interpreting the results of the analysis.

3.2.7 Chosen indicators – summary

In this chapter, we established a theoretical framework (three dimensions of tabloidization) in which we summarized available indicators of tabloidization. We then established our own criteria that helped us when choosing the indicators relevant for our research.

RANGE	FORM	STYLE
Number political or economic news (decrease)	Text length (decrease)	Negative sentiment of the article (increase)
	Number of images (increase)	Negative sentiment of the title (increase)
	Number of videos (increase)	Number of articles containing "!" or "?" in the title (increase)

	Image to text ratio (increase)	

Table 2: Chosen indicators – summary

3.3 ABOUT WALL STREET JOURNAL

Wall Street Journal is an American business-focused, English-language international daily newspaper based in New York city[8]. It has been founded in 1889 and it is a second largest US newspaper by circulation with 1,011,200 circulating issues (First place is occupied by USA Today with 1,621,091 issues)[9]. It belongs to **Dow Jones & Company** division of News Corp. The editorial pages of the WSJ are leaning towards American conservative (Bowden 2019). Most importantly for us, while it is published in a broadsheet format, WSJ is also published online, and the articles published online are also stored in a digital archive we can utilize for our research. WSJ also has its *PRO* section, offering a "premium suite of products for elite practitioners"[10] in a form of mostly finance content, which we will ignore for the purposes of our research. Uncharacteristically, WSJ converted its overseas printed version to a tabloid format on 17 October 2005. The change was however reverted in 2015.[11]

In 2016 WSJ switched to **digital-first** approach, splitting the newsroom into a digital desk and a print desk (Sterne 2016). This reorganization aimed to switch the focus mostly on mobile phone readership and was supposed to bring changes in terms of shorter, timelier, more digestible journalism to boost digital subscription in the rapidly evolving marketplace. The traditionally written format of the journal – as well as WSJ Pro subscription – was supposed to be kept only for relatively niche segment of the market of readers interested in content in such a form. The following paragraph is an excerpt from a memo sent by the Editor-in-Chief Gerard Baker (Mullin 2016).

[8] Source: www.wsj.com
[9] Source: www.cision.com/2019/01/top-ten-us-daily-newspapers/
[10] Source: https://pro.join.wsj.com/2015/08/15/is-this-a-question-4/
[11] Source: https://www.theguardian.com/media/2015/jun/11/wall-street-journal-european-and-asian-editions-broadsheet

For all reporters and editors, writing must come into sharper focus. We write many excellent stories, but in total, every day we write too many long stories and aren't nearly creative enough about how to tell stories in ways that engage our readers. We must urgently understand and address the reality that busy readers are looking to us to help them understand what is important and what not, what stories need a lot of time and focus and which ones less so. So we must be vigilant in keeping story lengths appropriate. Bluntly - but obviously, I hope - every story should be as short as it needs to be. There's no excuse for a single otiose word or punctuation mark in our writing. Too many stories have repetitive anecdotes or unnecessary quotes. We will cut them.

We selected WSJ for our analysis for several reasons. Firstly, WSJ has been historically adhering quite faithfully to our definition of broadsheet. Its main purpose was always to inform the readers about economic and political news and that has not changed throughout the years. As such, it represents almost textbook example of a broadsheet. Secondly, while digital archives are slowly becoming more and more common, WSJ archive is particularly well-made and very consistent. As such, it represents a perfect candidate for an analysis of such a character. Thirdly, while there has been a research targeting British media, where tabloidization has historically always been a part of the culture, the American newspapers represent much less explored territory. This, together with the fact that WSJ represents one of the largest journals in the world, offers an appealing research opportunity. Furthermore, the shift towards the digital-first type of format in 2016 together with the appeal for change of the internal values of the editorial board might represent a milestone that might just mark the change of the journalism practices of the newspaper. This event might prove to be a step towards more tabloidized version of the newspaper.

WSJ is also relatively consistent throughout the years. The journal has no obvious signs of a tabloid. You will not find any clickbait, provocative titles, or other obvious signs of tabloid in the newspaper or the web version of the journal.

Any indicators of tabloidization would be rather subtle. Qualitative analysis of the journal might not yield as conclusive of a result as a quantitative analysis trying to spot a subtle trend over a very large dataset. All the things considered; WSJ represents a very suitable candidate for a research of our character.

3.4 RESEARCH METHOD

To examine selected indicators, we decided to use quantitative analysis. We divided our process into two stages. For the first stage, we have created a web-scraping bot (**crawler**) that traverses through WSJ archive with the goal of collecting individual articles for our analysis. The logic of this crawler will be described in more detail further in the thesis. Besides the article content and all its parts (such as header, sub-header, images, etc.), the crawler also downloads selected *metadata* (such as word-count, number of images, article section, article keywords etc.,) that are present on the page. This metadata together with the article content form a basic data unit (article) that will be stored in a JSON format[12]. For every year in the period of 2000 – 2020 dozen of thousands of these data units will form a *year data set*. The total dataset will then consist of 756,889 articles stretched over the 21-year period.

For the second stage, we have created a program written in python (**analyzer**) – the logic of which will be described later within the individual analyses – that utilizes various programming libraries and modules to help us process individual indicators. The input for the analyzer consists of a year data set and the output is an average value for that data set for the defined parameters – e.g., the analyzer can tell us that the year *X* has an average *text length* of *Y* words. To see how a given parameter changed throughout the analyzed period, we will then compare the average for the first year (2000) with the average for the last year (2020). We can also compare the two periods of 2000 – 2010 and 2011 – 2020 to see the change between the two decades and the periods of 2000 – 2015 and 2016 – 2020 to compare the results before the WSJ implemented the digital-first policy with the period after. To try to showcase potential trends, we will also illustrate the

[12]JSON is a text format for storing and transporting data (source: www.w3schools.com)

data graphically. We have chosen a **run chart** as a most suitable means to graphically illustrate the results.

For each indicator we will present the following:

- A brief description of the logic of the analyzer in form of **pseudocode**
- The **relevance** of the parameter being analyzed in the context of tabloidization
- The expected **accuracy** of our analyzer
- The **run chart** displaying the trend of the analyzed parameter
- **Interpretation** of the results
- **Summary** table

3.5 RESEARCH QUESTIONS

Based on the chosen indicators, we will try to answer the following research questions in the context of the period of 2000 – 2020:

1. How did the article text length change?

The digital archive presents the author of the article with the benefit of not being limited by the physical constraints of the printed format. The author can simply utilize the concept of inverted pyramid – Placing the most fundamental information in the lead paragraph of the story, and then arranging the remaining details, from most important to least important (Scanlan 2008) to write freely. Furthermore, the digital nature of the articles allows the authors to edit or add additional content or audiovisual material to the article later, should the situation warrant it, so the circumstances and the limiting factors slightly differ from the printed format. However, the decreasing text length has been one of the most often quoted indicator of tabloidization and we do consider it to be fully relevant even in case of the digital format. Trend of a decreasing word count would signify less textual information presented to the user and could be connected to the limited depth of a journalism due to the economic factors even in case of digitally published articles. Therefore, we will consider a potential trend of a decreasing word count to be an indicator of tabloidization.

We can analyze year data sets to get average word count easily and accurately, as the word count is simply present in the metainformation of the articles during the crawling process.

2. How did the number of images within the article content change?

Increase in the amount of image material is another often-quoted indicator of tabloidization, but in case of digitally published media the conditions are different. While in case of the printed newspaper format the increase in space allocated to images logically results in decrease of space allocated to text (assuming the format stays the same), the same cannot be said for the digitally published articles. Due to the fact the author of the article has technically an unlimited amount of space at his disposal, the mere increase in the amount of space allocated to the image material might not necessarily mean limiting the amount of textual information to the reader. On the contrary, increase of the number of images might signal the increase of amount of information presented, assuming the amount of text stays the same. As such, we will not consider the potential increase in number of images to be a clear sign of tabloidization, unless the text length decreases.

As is the case of the word count, image count is another piece of information that is a part of the metadata present in every article. As such, this is another indicator that can be analyzed easily and accurately.

3. How did the number of videos change?

While video material arguably plays the same role as image material, and by the same logic could be considered indicator of tabloidization (especially in relation to the text), we will consider number of videos to be relatively irrelevant when it comes to indicating tabloidization. This is due to several factors. Firstly, we cannot programmatically detect whether the video is just graphic illustration or if it is a video report, therefore we cannot programmatically assess the "depth of journalism" when it comes to videos. Secondly, the presence of embedded videos on the web page is relatively recent trend that grew over the recent years with increasing capacity of bandwidth and technological capabilities of our mobile phones. As such, the growing trend of number of present videos is basically

unavoidable. We will still include this parameter in our analysis however, as it has some relevancy (sharply growing trend of number of videos in the last couple of years might signal that the journal relies more and more on the video material), it is easy to add, and can serve as a basis for further research.

The number of videos must be captured programmatically, as it is not present in the article metadata. We can easily traverse the article body and detect the number of videos (iframes) present.

4. How did the image to text ratio surface ratio change?

While the mere number of images may not be a relevant indicator of tabloidization by itself in the digital format, the decrease of space allocated to the text at the cost of increase of space allocated is arguably more relevant. As such, we will look for a trend of decreased space allocated to text and increase of the space allocated to images.

To perform analysis of such a kind, we will have to pull the relevant data programmatically. Our crawler will load the images and capture their dimensions (width and height). It will then capture the same dimensions of the entire content of the article. We can then calculate the surface dedicated to the article and what percentage of it is dedicated to images. Note that for this analysis, we will focus merely on the article content (the main body of the article) and ignore the area dedicated to the cover image[13] and header. The reasoning for this is two-fold – firstly, this area can take on many forms that changes throughout the years, e.g., no cover image, a simple image, large cover image, a layout covering entire page, etc. To try to programmatically capture the area could prove inconsistent and unreliable, which could skew the results of the analysis, should the banner area be included. Secondly, we do not consider the area to be a convincing indicator of tabloidization. WSJ kept adding more image material to the article header, as we can see from no cover images in the year 2000 and all the articles having some sort of a cover or introductory image in 2020, but there has been a general trend in web development to gradually move to more and more graphic user interfaces throughout the years as the bandwidth capacity became less of a limiting factor. Furthermore, the

[13] Large images in the header area that serve the purpose of introducing the visitor to the article

presence of the cover images might not necessarily come at the cost of the amount of text, so we will ignore this area for our analysis.

5. How did the negative sentiment change?

While the sentiment might generally oscillate based on many variables (e.g., current political or economic situation), many authors consider increase in negativity growing trend of tabloidization (Norris 2001, Jelínková 2019). We will therefore treat potentially increasing trend of negative sentiment as a reliable trend of tabloidization.

It is possible to detect a sentiment with the use of VADER sentiment analysis tool described to the greater detail in chapter 3.7. The tool takes a text as an input and outputs negativity, positivity, neutrality score ranging from 0 to 1. It also outputs a compound score (on the axis where -1 is the most negative, 0 signaling neutral and +1 being positive). The tool is trained with the help of machine learning and is originally meant primarily for social media texts, but it is accurate enough to be helpful in other areas as well, including our research. We can then take the average negativity score for each year to spot potential trend.

6. How did the difference in sentiment of the title and content change?

We have mentioned that increased negativity correlates with user engagement and draws to online articles much needed visitors. Furthermore, we established that media are very often presented only as a combination of a thumbnail, title, and lead in the concurrent fast-pace world, especially on mobile devices, and last but not least, we introduced the notion of *clickbaits*, which exploit a completely different sentiment of the title (usually very negative) from the sentiment of the article (usually less negative) that became popular due to this logic, the difference between the sentiment of the title and content represent potentially powerful indicator of tabloidization.

For this analysis to take place we can utilize similar approach as in the RQ6, simply using the VADER tool to compare the sentiment of the title with the sentiment of the text to determine the difference between the two. Increasing difference between the sentiment of the title (being more negative) than sentiment

of the article content could then likely be considered as an indicator of tabloidization.

7. How did the count of punctuation symbols (?!) in the title change?

Given the importance of the title we will consider the presence of the question mark and exclamation mark to be a potential indicator of tabloidization, as the question mark itself signifies a question while the exclamation mark is often used to capture reader's attention.

Detecting the presence of these symbols is extremely simple and accurate from the programmatical point of view. We will then count the percentage of articles that do contain a question mark or an exclamation mark. Increased percentage of articles containing these punctuation marks will then be treated as a sign of tabloidization.

8. How did the percentage of political / economic news change?

The main goal of WSJ is to provide news related to the market from political and economic sphere, therefore we expect high percentage of hard news in the journal. WSJ also offers many articles not falling under political or economic category such as *Sports*, *Books & Arts* or *Life & Work*. Observing the number of hard news could serve us as a good indicator of tabloidization, should the number of hard news decrease over the years.

Calculating the number of hard news should not be particularly problematic. Since every article falls into certain sections (e.g., World, Home, Politics, Economy, Business, etc.) we can just calculate the number of articles not belonging to Politics or Economy over the years. Potential trend of decreasing number of hard news over the years would then serve as an indicator of tabloidization.

3.6 SIMILAR RESEARCH OVERVIEW

3.6.1 Research related to tabloidization

Although the term tabloidization itself is relatively new term (being used since the 1990s), there have been much research dedicated to this phenomenon. McLachlan and Golding (2000) conducted extensive analysis of tabloidization within the British national tabloid Daily Mirror and concluded it had decreased its text content (from 320 to 160 words in 1982 – 1997) and increased its picture content (increasing number of pictures per page in 1970 – 1992). This study set the theoretical framework based on which another quantitative content analysis was carried out in Britain, focusing on The Sun and The Mirror (Audit Bureau of Circulations, 2001) over 10-year period. The study concluded the decrease of space per page devoted to information at the cost of services and entertainment, increase in foreign affairs news stories compared to home news and increase in visual elements at the cost of text. Esser (1999) conducted an analysis of the press of Britain, Germany, and US, concluding that German readers don't appreciate tabloids as much as British readers, and that trends towards tabloidization are much less detectable in Germany, stating that the amount of competition on the market is a decisive factor for the progress of tabloidization (318).

As for the theses dedicated to tabloidization, In the Czech Republic, Jelínková (2019) was interested in quantitative analysis of tabloidization of journals *Mladá fronta Dnes* and *Blesk* in the period of 2005 – 2017. While she detects a small amount of tabloidization, only 2 out of her original 10 hypotheses got confirmed – increase in the number of colorful headers and space dedicated to image material. Zapletalik (2020) explored the influence of tabloidization on Sport News in the Czech Republic and USA but found merely a small evidence of tabloidization regarding the style dimension, not so much regarding the soft news / hard news ratio. From outside of the Czech Republic, Brandelid and Eklund explored signs of tabloidization in Swedish newspapers regarding the coverage of COVID-19 pandemic. They managed to find just a minor evidence of tabloidization which they attributed towards increase in competition between the Swedish journals. The result of their study also goes hand in hand with the importance of social context of the tabloidization, as it indeed seems that certain nations simply

tend to accept tabloids much less (e.g., Germany and Scandinavia) than others (UK).

We can see that results of studies dedicated to tabloidization historically led only to partial confirmation of the phenomenon with a predominant trend of large percentage of hypothesis set by researchers remaining unconfirmed. Rowe reached a similar conclusion, claiming that "tabloidization is happening in the general and that it is a clear feature of developments in the broadsheets, but that it is not occurring consistently, and that in some areas the tabloids may be moving in a different or less predictable direction" (2009, 356).

3.6.2 Research related to computer analysis of natural language

As for the computer analysis part of our research, a research of a similar nature has been conducted by data analyst Phillipe Heitzmann (2020). To investigate a possible relationship between article emotionality, subjectivity, positivity/negativity, and user engagement (number of comments). In his analysis of 22,772 WSJ articles, Heitzmann utilized *selenium* for data collection process and *VADER* analyzer for sentiment analysis – tools that form a part of our toolkit as well – and managed to find a positive correlation between negative sentiment of the article and number of comments (and therefore page visits). His work indirectly confirmed that news of negative sentiment led to bigger user engagement, although his research was not concerned with the reasoning of this phenomenon.

Alison Salerno (2020) utilized various machine learning algorithms to try to develop a tool that would be able to analyze whether a header is or is not a clickbait. The result was a tool with 90 – 93% accuracy.

Studies like these present a convincing argument that computer analysis of natural language has its place and usability in the linguistics field, as it seems that tools utilizing machine learning seem to generally lead to conclusive results.

3.7 TECHNOLOGY USED

While the nature of this thesis is predominantly linguistic, we believe if we did not at least briefly introduce utilized technology and frameworks, we would be doing a disservice to the authors of the used tools, a reader interested in the technical aspects, or somebody who would like to conduct a research of a similar nature on their own. With that in mind, let us take a brief look on the programming tools utilized in our thesis.

Python

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation (Kuhlman 2012). While the choice of the programming language is perhaps not as crucial as the choice of the tools themselves, its simple syntax, versatility, and large community makes it a great tool to use for almost any project. Furthermore, python is an amazing tool for handling web, AI and machine learning, and data analysis – all the issues we will have to deal with during our data collection and data analysis processes. Perhaps most importantly, python allows us to utilize **Natural Language Toolkit** (NTLK), which is an important part of our analysis process which we will introduce further in this chapter.

Selenium & ChromeDriver:

Selenium is an open-source automated testing framework for web applications (Rungta 2021). Selenium becomes a powerful tool when combined with ChromeDriver – "an open-source tool for automated testing of webapps across many browsers. It provides capabilities for navigating to web pages, user input, JavaScript execution, and more"[14]. In its essence, selenium is just a testing framework. Its task is to automate simple actions that a human could perform to make sure the application works correctly – e.g., *open a Chrome web browser, visit a webpage A, make sure that an action B takes place if we click a button C*.

[14] Source: https://chromedriver.chromium.org/

Selenium can also be utilized for slightly more complex, repetitive tasks, and we will use it for the data collection process. While Selenium is arguably not the fastest, nor the most efficient tool for such a task (compared to e.g., *Scrapy*[15]), it is relatively easy to use, represents a low workload for the WSJ servers, and bypasses some precautions that WSJ is taking to protect themselves against web-scraping.

Natural Language Toolkit (NTLK) & VADER Sentiment Analyzer:

NTLK is a free, open-source, community driven project that represents "a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum […] NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."[16]

While NTLK represents an amazing tool and one of the best libraries for programmatically processing natural language, we will be mostly interested in its sentiment analysis tool – **VADER.** VADER (**V**alence **A**ware **D**ictionary for s**E**ntiment **R**easoning) …is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. It is available in the NLTK package and can be applied directly to unlabeled text data" (Beri 2020). In its essence, we can run a text through VADER and as an output we will receive an information whether the sentiment of the text is positive, negative, or neutral. Although Vader was primarily developed for social media texts, according to its developers, it is also applicable to other domains[17], and it should be sufficient for our purposes as well.

---

[15] Scrapy is an application framework for crawling web sites and extracting structured data written in Python (source: https://doc.scrapy.org/en/latest/intro/overview.html).

[16] Source: https://www.nltk.org/

[17] Source: https://github.com/cjhutto/vaderSentiment

4 COMPUTER ANALYSIS

The following chapter will be interested in data collection and subsequent data analysis. To familiarize the reader with the underlying processes taking place, each process will be briefly introduced with a short pseudo-code section that will shortly describe the core logic of the program without requiring any technical knowledge on the side of the reader.

Each process will have its flaws and imperfections, which arguably might negatively impact the outcome of the thesis. Furthermore, each process has a different degree of accuracy. Some will be extremely accurate (e.g., word count analysis should be nearing 100% accuracy), and some arguably inaccurate (VADER analysis, assuming near 60% accuracy). These imperfections can be mostly fixed and further optimized given enough time and resources but given the nature of our research and the sample size of analyzed articles, we consider these flaws and inaccuracies acceptable, as the variance in results should be negated by a large sample size. Therefore, while we will do our best to describe potential flaws and shortcomings of each process, it is important to keep in mind that we are striving to prove / disprove a trend rather than to develop a perfectly optimized process.

4.1 DATA COLLECTION PROCESS

Pseudocode

```
# go to WSJ archive at https://www.wsj.com/news/archive/years
# login with credentials
# for each year store a link for each month
# # for each month store a link for each day
# # # for each day store all article links
# # # if the day has multiple pages store those links as well
# # # # for each article link:
# # # # if the article contains purely audiovisual material - most likely a slideshow, and does not contain article text, mark it as audiovisual and don't do anything.
# # # # Otherwise get the following: header, URL, sub-header, full article content (body of the article), description, webpage section,
```

```
article type, article keywords, article content dimensions (width
and height), word count, image count, header image dimensions,
article content image count, article content image sizes,
audiovisual flag.
```

For the collection process we opted to use **Selenium + ChromeDriver**. The crawler will visit the WSJ **archive** to try to scrape every article in 2000 – 2020 period and collect the data required for our analysis. WSJ stores a certain amount of data in the webpage metadata for various purposes. This is helpful to get to certain information directly (e.g., article word count). Other data we must gather programmatically (e.g., image sizes and content size for our analysis of *image to text ratio*.

WSJ produces large number of articles on daily basis – roughly between 20 and 300. During the scraping process, we have decided to ignore all the articles falling under *WSJ Pro* subscription, which eliminates a small percentage of the article pool from the data set. Most of the articles follow the same pattern – a typical article will contain a header (often containing a title and a cover image), sub-header (containing a lead), and an article content (containing text, possibly several images, occasionally an embedded video).

However, not all the articles are of the same nature. Some articles deviate slightly from the traditional formula and lack a cover image, or description, or other elements. Other deviate significantly – some are purely audiovisual – e.g., a video, or a slideshow without any text content outside of the audiovisual iframe[18]. Some articles might be just a simple image. Finally, certain number of articles seem borderline unique or are exceptionally different – e.g., interactive tour through a plane crash utilizing 3D imaging. Overall, WSJ articles differ significantly, and WSJ is not afraid to experiment and offer large variety of material, which represents an obstacle when retrieving the data.

While it is theoretically possible to capture information these articles programmatically, we will be mostly interested in the articles that follow the standard formula, and overwhelming majority of the articles will fall into this category. In case we fail to retrieve a text content from the article, and detect a slideshow, we mark the article as an **audiovisual**. Besides WSJ Pro articles, we will also ignore all the articles that deviate significantly from the standard formula, as

[18] Interactive frame used to embed another document within the html document (frequently video).

our crawler would not process them correctly. Small percentage of articles might fail due an error on the side of WSJ, as some articles might return a HTTP error 404[19], signaling that the article is no longer present in the archive, and some might return HTTP error 503[20], signaling that the WSJ server is currently unable to process the request (likely due to the heavy traffic period). Finally, a very small number of articles might fail due to the connection error on our part.

While we ignore certain percentage of articles (the biggest number consists of the WSJ Pro articles), the effect of the missing data on the conclusion of the analysis is arguably negligible. For the thesis, our goal was to gather at least 10,000 articles per year, totaling 250,000 articles. We eventually managed to gather 756,889 articles complete number, averaging 36,042 articles per year, which should represent a sufficient sample size for our research. Note that for the purpose of our research we will distinguish two types of articles:

- **audiovisual articles**: usually does contain a slideshow, image material, video material, or something completely different. This type of the article does not contain the traditional article body and often varies from the standard article layout. It is difficult to process such an article programmatically in a consistent manner. These articles form a small percentage of the analyzed material and will be frequently omitted due to these reasons.
- standard **article**: This article follows the traditional formula of journalism. For the purposes of the thesis, we will divide this article into:
    - **article header**: This section usually contains a headline, byline, a picture, cover image or a banner, which were generally not present in the earlier years of our dataset. As was the case of audio-visual articles, this area also varies. It can consist of an overlay taking over entire page, slideshow, or might be missing entirely. Because of these reasons, this area will also be omitted from several parts of our analysis.

[19] Returned when the page or item is not found
[20] Returned when the service is currently unavailable

- **article body:** This section will contain the main content (text) of the article, often accompanied by audio-visual material – most frequently images and videos in the recent years

4.1.1 Data entry (article) structure

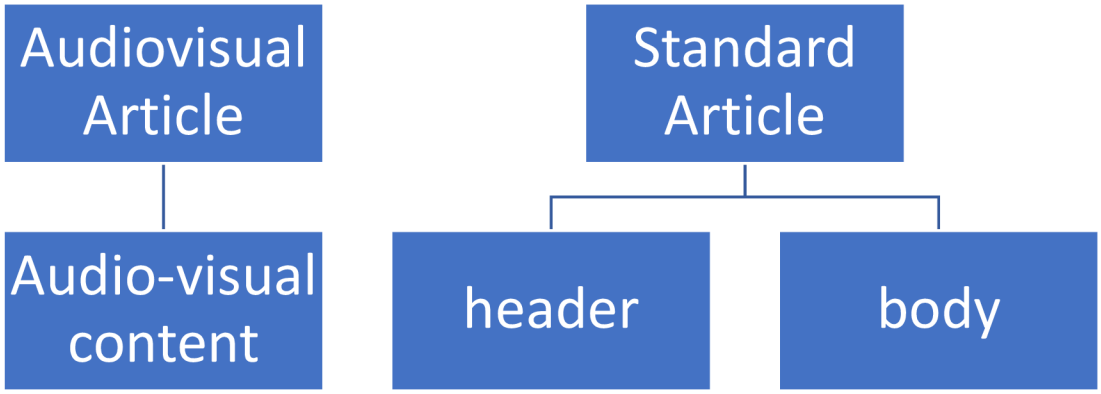

Figure 1: Data entry (article) structure

4.1.2 Dataset summary

Year	Total Articles	Audiovisuals	Audiovisuals (%)
2000	51414	0	0
2001	42836	0	0
2002	23686	0	0
2003	38916	0	0
2004	21849	2	0.009
2005	40156	135	0.34
2006	42441	350	0.83
2007	24004	421	1.75
2008	51593	1179	2.29
2009	27377	964	3.52
2010	26971	1112	4.12
2011	19045	903	4.74
2012	43247	2304	5.33

2013	64263	2872	4.47
2014	39376	1388	3.58
2015	43002	1545	3.59
2016	20878	534	2.56
2017	44136	998	2.26
2018	37748	925	2.45
2019	18237	286	1.57
2020	35779	424	1.19
AVG	**36042**	**778**	**2.12**
TOTAL	**756889**	**16342**	X

Table 3: Collected dataset summary

4.2 Q1: HOW DID THE ARTICLE TEXT-LENGTH CHANGE?

Pseudocode

```
# for each data entry:
# # get the article word count obtained from the article metadata
# # if the article is audiovisual, treat the word count as 0
# # get the average word count per year
```

Relevance

As established in the chapter 3.2, the text length is one of the of the prominent indicators of the *form* dimension. While in the printed format of the newspaper the decreasing text length usually signifies more room being allocated to the image material, in digital format the parameter might signal generally shorter articles, potentially less textual information presented to the user, quantity of the articles at the cost at quality and overall, more *shallow* form of a journalism. As such, we consider the decreasing text length an indicator with **high relevancy**.

Accuracy

Since this parameter is present in the WSJ metainformation we consider our method to be **highly accurate**. While a mistake on the side of the WSJ is possible, we tested the parameter to make sure the actual word count matches the word count stated by the WSJ. Marking the audiovisual articles as articles with a word count of 0 for the purpose of this analysis arguably skews the results to a certain extent, as even audiovisual material will have titles, descriptions, speech in videos etc.

To achieve fully accurate method, the audiovisual articles would probably have to be taken into the account, it is however important to note that the number of audiovisual articles is extremely small to cause any significant difference in the results, and accounting these articles for 0 words is arguably an acceptable solution

Run Chart

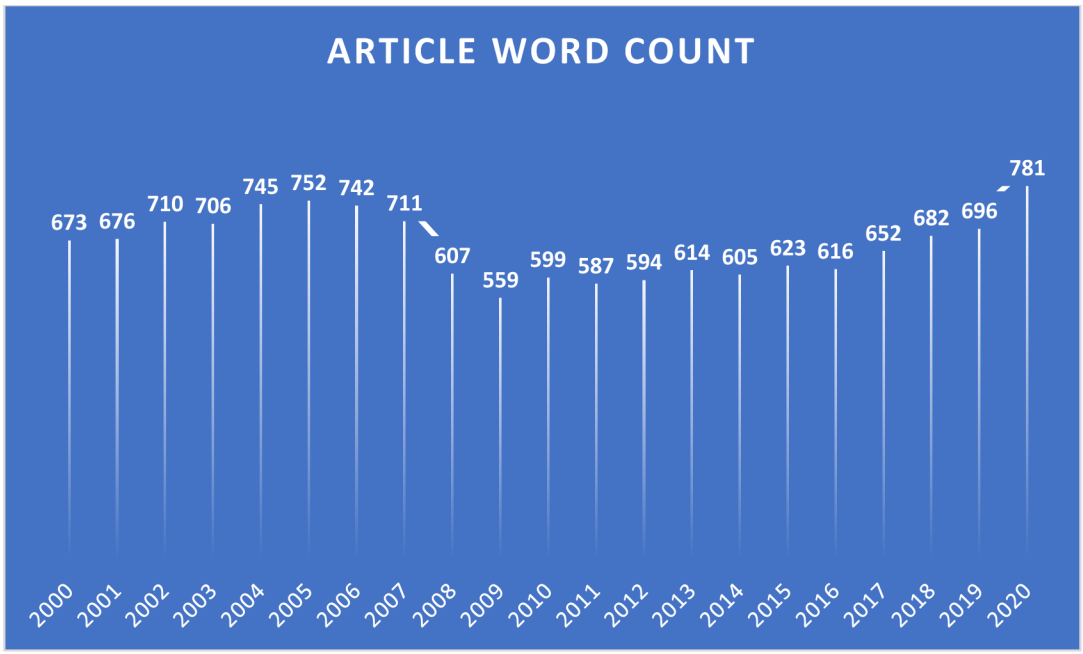

Figure 2: Article word count (value represents average word count in article content)

Results Interpretation

From the Figure 2 we can see that the article word count remains relatively constant throughout the years, with the lowest value of 559 words per article in 2009 and highest value of 781 per article in 2020 This revelation is perhaps surprising,

especially given the Baker's statement about focusing on shorter, more relevant articles and elimination of "repetitive anecdotes or unnecessary quotes" (Mullin 2016).

Not only the word count does not decrease throughout the years, but the peak year with the highest word count is the year 2020 – the most recent year of our analysis. The difference between the first year and the first year is increase of 108 words (16% increase), which signals a small increase of the word count throughout the years. The period of 2000 – 2010 then averages 680 words when compared to the 645 average words of the period 2011 – 2000, which marks 35 words (5.2%) decrease between the two decades. This, however, especially combined with the first year / last year difference cannot be considered an indicator of tabloidization. The period of 2000 –2015 averages 657 words when compared to 685 words in 2016 – 2020, marking 4% increase, despite the expectation of lower word count in this period.

Summary

Article word count	
2020 vs 2000	***16% increase***
2000 – 2010 vs 2011 – 2020	***5% decrease***
2000 – 2015 vs 2016 – 2020	***4% increase***
Accuracy	***High***
Relevance	***High***
Trend of tabloidization	***No***

Table 4: Article word count – summary

4.3 Q2: HOW DID THE NUMBER OF IMAGES WITHIN THE ARTICLE CONTENT CHANGE?

Pseudocode

```
# for each data entry:
```

```
# # get the article content image obtained during the scraping process by calculating the image elements within the article content
# # if the article is audiovisual, skip the data entry
# # get the average article content image count per year
```

Relevance

While the increase of image material is considered to be relatively good indicator of tabloidization in the written newspaper, where it usually signifies the decrease amount of text (unless the newspaper's format changed as well), we cannot give the same relevancy to this indicator in case of digitally published media, where amount of images might simply provide further information while the amount of text remains the same, as the digital format does not face the same space constraints as the printed format. This – combined with the fact that according to our answer to the research question #1 the amount of text did not change – means we will consider the number of images within the article content to be an indicator of **low relevancy.**

Accuracy

As was the case of the *article word count*, even this parameter is present in the WSJ metainformation. However, the parameter value in the article metainformation seems to be often incorrect, or completely missing. As a result, we retrieved the parameter programmatically by simply counting all the image elements in the article content. We then tested the number by comparing it to the image count parameter from the article metainformation (in cases where it was present). Based on our results we consider the data to be almost 100% precise and this method to be **highly accurate**.

It is important to note, however, that the header of the article was skipped for this analysis, which frequently hosts a cover image or an image layout. The reasons for not including this part were discussed in the chapter 4.1, and inclusion of the header would almost definitely lead to a positive result (trend of growing number of images), as the cover image was not a common part of the article in the first years of our data set but became one over time. For our purposes, we decided

to analyze purely the area of the article where image material "competes" with the text.

As in the Q1, the audiovisual articles were skipped for this analysis, as we do not have a precise method of pulling the parameter out of those articles. Again, the impact of us skipping these articles should be negligible.

Run Chart

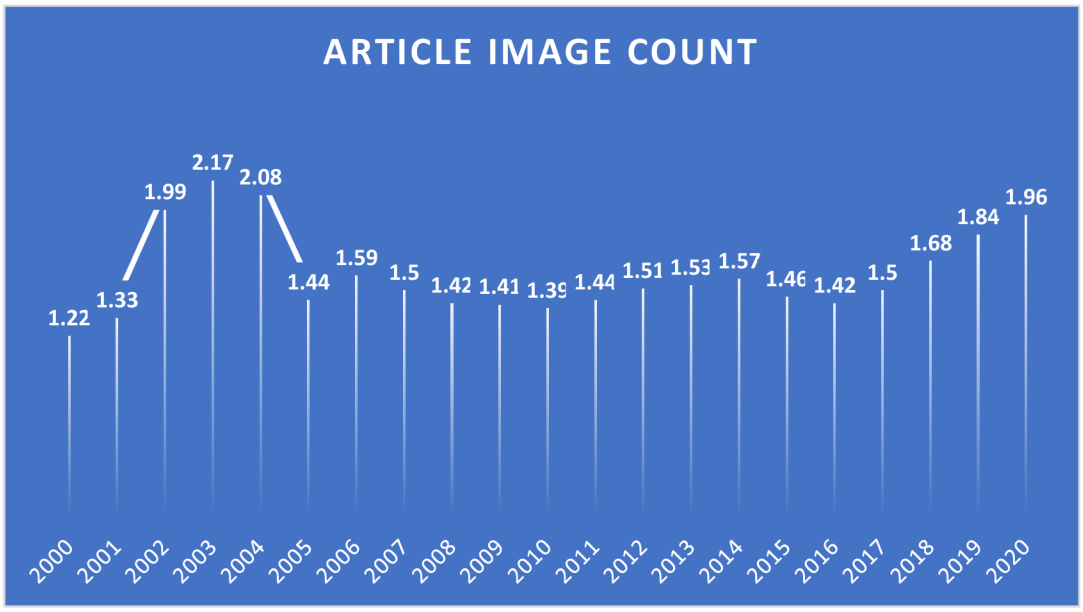

Figure 3: Article image count (value represents average number of images per article)

Results Interpretation

From the Figure 3 we can see that the image count is indeed lowest in 2000, surprisingly followed by a sharp increase in the period of 2002 – 2004. The year 2003 then represents a peak with 2.17 images per article in 2003. The spike then corrects itself and remains relatively steadily close to the average 1.56 images in the following period up until the year 2016, where it starts increasing again, with second peak at the most recent year 2020 with 1.96 images per article.

Overall, the trend of the number of images per article goes according to expectations, with the lowest year being 2000 and the (second) highest year being 2020, marking 60.5% increase. We can also observe the growing trend of image material beginning in the year 2016. The extraordinary peak in 2002 – 2004

however results into both 2000 – 2010 and 2011 – 2020 periods having the same number of images per article (1.59). The period before 2016 then shows 1.57 articles when compared to the slight increase to 1.68 images in 2016 – 2020, marking a mere 7% increase.

Summary

Article image count	
2020 vs 2000	***60.5% increase***
2000 – 2010 vs 2011 – 2020	***0% increase***
2000 – 2015 vs 2016 – 2020	***7% increase***
Accuracy	***High***
Relevance	***Low***
Trend of tabloidization	***No (Growing trend after 2015)***

Table 5: Article image count – summary

4.4 Q3: HOW DID THE NUMBER OF VIDEOS CHANGE?

Pseudocode

```
# for each data entry:
# # get the number of videos within the article content obtained with the crawler
# # get the average number of videos within the article content per year
# # if the article is purely audiovisual, skip the article from the analysis
```

Relevance

Like in the case of number of images indicator, we consider the number of videos to be of **low relevancy**. While increased reliance of the audiovisual material generally signals the trend tabloidization, we cannot programmatically reliably

analyze the length or the content of the videos. Furthermore, since the videos have not been used in the beginning of the period of 2000-2020, and even after their emergence as a part of newspaper reporting, their popularity was growing in correlation with bandwidth limitation becoming less and less prevalent over time and growing popularity of video hosting services such as YouTube. We can thus almost definitely expect positive trend to be found during our analysis. Yet, we chose to keep the analysis present due to its informative function. It can also show how much WSJ depends on videos in their online journalism, which can serve as a basis for a further research. We can also look for any "drastic" jumps in video number to try to spot any severe increase in the use of video material.

Accuracy

Video count that we had to process programmatically as it was not present in the WSJ metainformation. As we simply programmatically counted the video count in the body, and completely omitted the article header, as a part which does not tend to have any videos present, we consider this method to **be highly accurate** (almost 100% precise).

Note however, that we did decide to omit purely audiovisual articles from our analysis, as counting the video elements is not easily done in a consistent manner. As usual, given the small amount of the audiovisual articles, the impact on the result should be negligible.

Run Chart

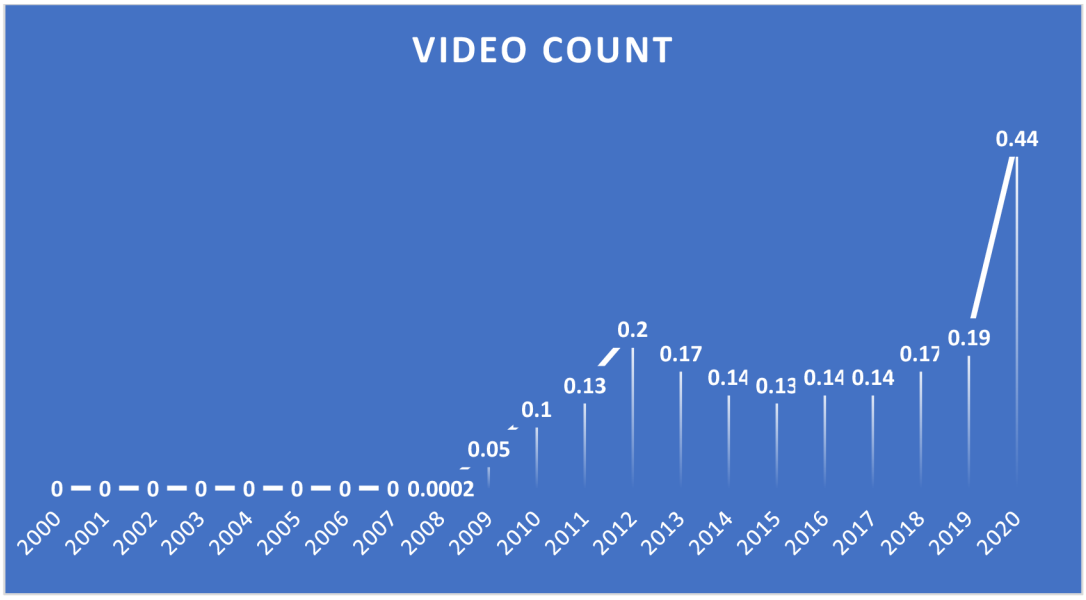

Figure 4: Video count (value represents average number of videos per article)

Result interpretation

From the Figure 4 we can indeed see the that videos were not really used in the articles until the year 2008. Because of that, we can objectively disregard the time periods of 2000 – 2010 and 2000 – 2016 from our analysis. We can observe that the videos have been a common part of the WSJ journalism in 2010 – 2020, averaging 0.185 videos per article, or 1 video every 5.5 articles. As mentioned, we cannot really interpret this phenomenon as an indicator of tabloidization.

The sharp increase between the years 2019 and 2020, however, is at the very least noteworthy. The increase from 0.19 videos per article (or 1 video in 5 articles) in 2019 to 0.44 videos per article (almost one video every two articles) represents a 132% growth on year-to-year basis, and while there is always a possibility of a statistical deviation, our dataset consists of 18,237 articles for the year 2019 and 35,779 articles for the year 2020. While we cannot really consider *number of videos* as a reliable indicator of tabloidization, this signals significantly increased reliance of WSJ on videos in the recent years.

Summary

Video count	
2020 vs 2000	*X*
2000-2010 vs 2011-2020	*X*
2000-2015 vs 2016-2020	*X*
Accuracy	***High***
Relevance	***Low***
Trend of tabloidization	***No***

Table 6: Video count – summary

4.5 Q4: HOW DID THE IMAGE TO TEXT SURFACE RATIO CHANGE?

Pseudocode

```
# for each data entry:
# # get dimensions of all images and videos within the article content(width and height)
# # get dimensions of the entire article content (width and height)
# # calculate total surface area occupied by the article content
# # calculate total surface area occupied by the audiovisuals
# # get the average percentage of surface area occupied by images per year
# # if the article is purely audiovisual, the area value is 100%
```

Relevance

While we gave the *number of images* and *number of videos* low relevancy, as in case of digital media it might not necessarily signal the decrease of the text, the same cannot be said about the *image to text surface ratio*. Increase in the surface allocated to audiovisuals means less area dedicated to text and more reliance on the audiovisuals. Note that this question might not necessarily correspond to the number of images in the text, as various images can have different dimensions and

play slightly different roles in the article body (e.g., illustration image vs whole page overlay incorporated in the image). As increase of the surface area allocated to images (audiovisuals) and decrease of the area allocated to the text is one of the predominant tabloidization indicators we discussed in the theoretical part, we consider this indicator to be of **high relevancy.**

Accuracy

While this parameter is not present in the article metadata, we can get the image and video dimensions accurately and reliably with the help of our selenium *crawler* during the data collection process. Even though we are detecting all video and image elements in the article content, it is technically possible we might miss other audiovisuals (e.g., an audio track embedded in the article would pass unnoticed through our crawler). Despite that, we assume this method to be **highly accurate** nearing 100% precision, as the number of potential non-video, non-image elements present in the WSJ archive is close to zero. Note that pure audiovisual articles are included in this part of the analysis and will be simply considered to have 100% of their area occupied by audiovisuals.

As with previous research questions, we left out the article header from the analysis, the reason again being that this area is difficult to programmatically process in a consistent manner, as well as the fact that including this area would basically guarantee the chance of positive trend occurring.

Since the number of videos will arguably skew the resulting data and increase the chance of the positive trend findings, we will also provide a result without inclusion of videos in the process. This will however serve mostly as a point of reference, and the videos will be included in the result interpretation.

Run Chart

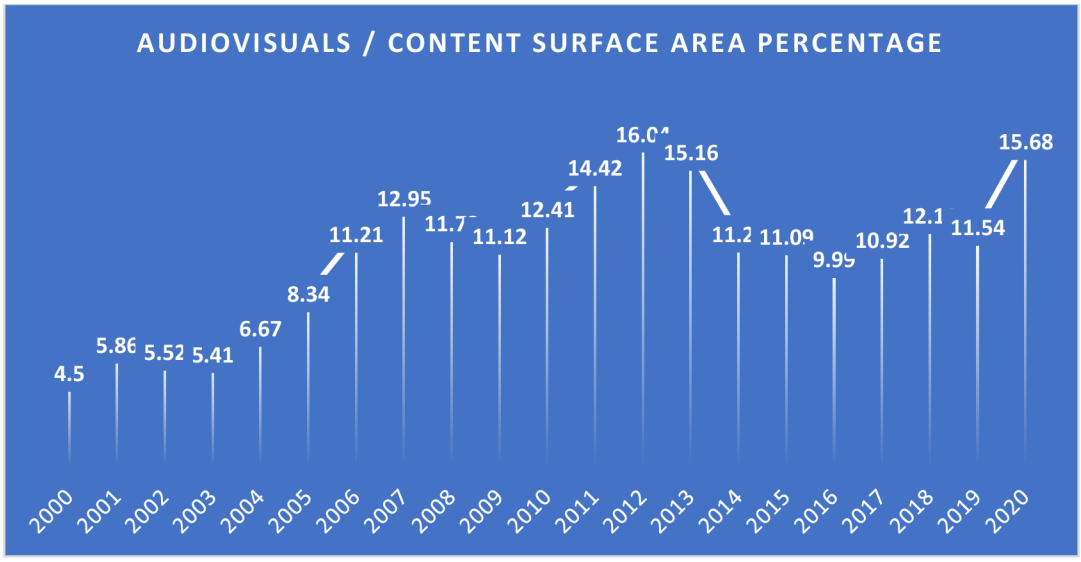

Figure 5: Audiovisuals to text ratio (value represents percentage of area occupied by audiovisuals)

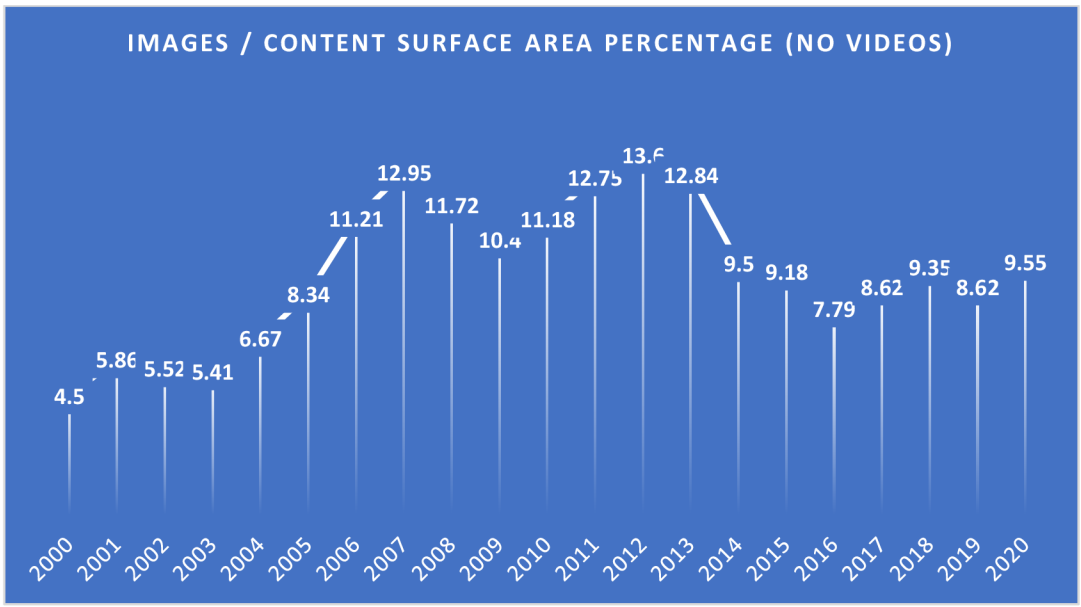

Figure 6: Images to text ratio (value represents percentage of area occupied by images without videos)

Results Interpretation

From the Figure 5 we can clearly see the growing trend of area occupied by audiovisual material throughout the years. The year with the lowest area occupied by audiovisual material is indeed year 2000 with 4.5% of area occupied by images while the most recent year marks the year with the second highest (15.68%) area occupied by images and videos. This marks 248% difference between the first and

last year. Surprisingly, we can find the lowest area dedicated to the text during the years 2011 – 2013, with 2012 being the peak with 16.04%, and not in the final years, as we might expect.

Overall, there is a clear difference between the periods of 2000 – 2010 (avg 8.7%) and 2011 – 2020 (avg 12.8%) marking 47% increase in the area allocated to the audiovisuals. While there is a 17% increase in the periods of 2000 – 2015 and 2016 –2020, this can be hardly attributed to the change in the company approach, as it rather seems to be a part of a growing trend over the years.

While we can clearly see a similar trend even if we exclude the videos from the analysis, the peak in the final years becomes less prevalent, as we can clearly see the impact of video material as described in the Q3.

Summary

Audiovisual surface area	
2020 vs 2000	***248% increase***
2000 – 2010 vs 2011 – 2020	***47% increase***
2000 – 2015 vs 2016 – 2020	***17% increase***
Accuracy	***High***
Relevance	***High***
Trend of tabloidization	***Yes***

Table 7: Audiovisuals to text ratio – summary

4.6 Q5. HOW DID THE NEGATIVE SENTIMENT CHANGE?

Tool Introduction

For the next two analyses we will utilize the VADER sentiment analysis tool we briefly introduced in the chapter 3.7. To give a better perspective about how the tool works: "VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text" (2020).

In summary, VADER divides the text into tokens (words), rates each token on a sentence level and produces a compound score for the analyzed text. The score is a float between 0 – 1 for positive, neutral, and negative sentiment (0 being the least of a given sentiment, 1 the most). VADER also produces a compound score in a form of a float between -1 and 1 (-1 being negative, 0 neutral and 1 positive sentiment). We can therefore say that 1 point of rating represents a difference between positive and neutral or neutral and negative sentiment. Figure 7 illustrates the analysis on an example, where the sentence represents the input, and the scores represent the output of the analyzer. Notice that VADER is capable of distinguishing even some of the tricky scenarios – *This is shit.* is rated -0.5563 (negative) while *This is the shit!* is rated 0.6476 (positive).

For the sake of simplicity, we chose to use the compound score for the purpose of this analysis. The compound will allow us to not only see any trends in the article sentiment, but it will also allow us to see whether the article is leaning more towards positive or negative sentiment.

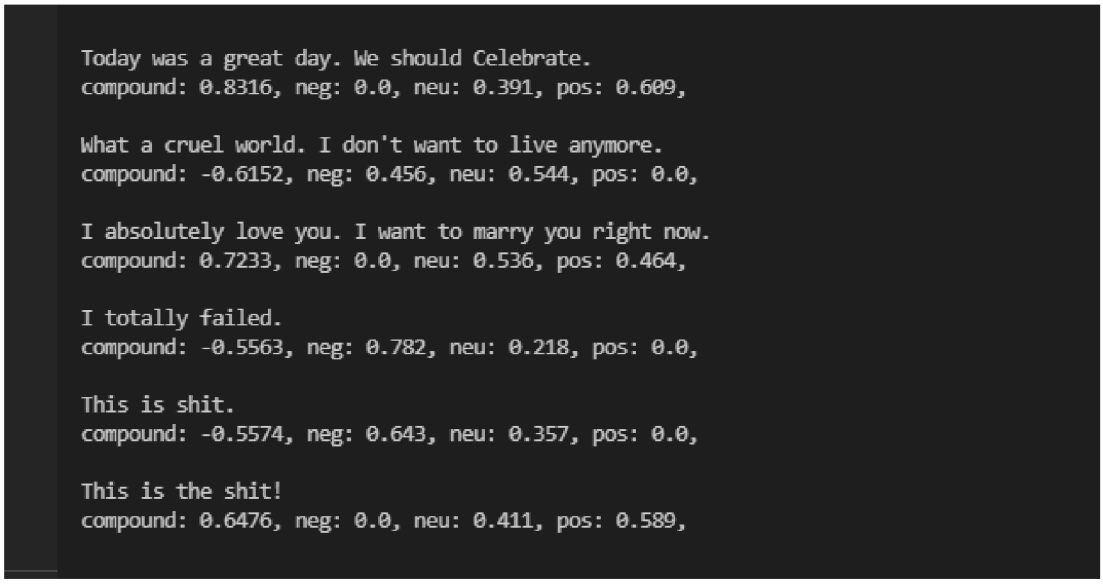

Figure 7: VADER analysis example

Pseudocode

```
# setup the VADER analysis tool
# train the VADER with the native provided corpora
```

```
# for each data entry:
# # if the article is purely audiovisual, skip the article from the analysis
# # process the article text with the VADER analysis tool
# # do not include header and article lead
# # get the compound average per year
```

Relevance

While negativity is an often-cited indicator of tabloids, it is not an indicator that would receive a lot of attention during research focused on tabloidization. We can attribute this to the fact that from the linguistic point of view it is a very difficult to grasp negativity in any objective way, as it necessitates a large degree of subjectivity or a very high-quality framework for analysis. Furthermore, sentiment is arguably always subjected to number of variables, such as seasonal events (elections, flu season), economic state (depression, inflation), or global and local events (pandemic, crime increase), which introduces variance when dealing with limited dataset or a research of a qualitative research. From the programmatical point of view, research of the similar nature is quite frequent since the emergence of natural language analysis tools such as NLTK and sentiments analyzers such as VADER or TextBlob as the tool is very consistent in the sentiment evaluation and any variance in the score should even out over long term and sufficient sample size, as proved by studies of Heitzmann (2020) and Salerno (2020), so we consider the indicator analysis to be perfectly viable.

Despite its low presence in linguistic sphere when analyzing tabloidization and aforementioned complications, since Heitzmann managed to prove that negativity is linked to the user engagement and the number of visits within WSJ (thus that increase in negativity in theory leads to more profit), we consider negativity to be an indicator of **high relevancy**, especially in the digital environment.

Accuracy

Before processing the text with the VADER tool, we tried to process 10 data entries to see whether the tool is reasonably accurate in case of WSJ articles. The results were surprisingly decent on the small sample, 8 articles were evaluated reasonably correctly, 1 was incorrect (deviation less than 1) and 1 heavily incorrect (deviation more than 1). Despite this, we consider the question Q5 (and Q6) to be the least accurate part of our analysis and will assign them **low accuracy**. According to its developers, the tool can be considered 60% accurate, and we will expect similar, or perhaps even a slightly lower performance in our case as well. The relative inaccuracy of this tool should be however greatly diminished by the very large sample size and a potential trend should successfully emerge. As in other partzs of our analysis, the audiovisual articles will be omitted from this process.

Run Chart

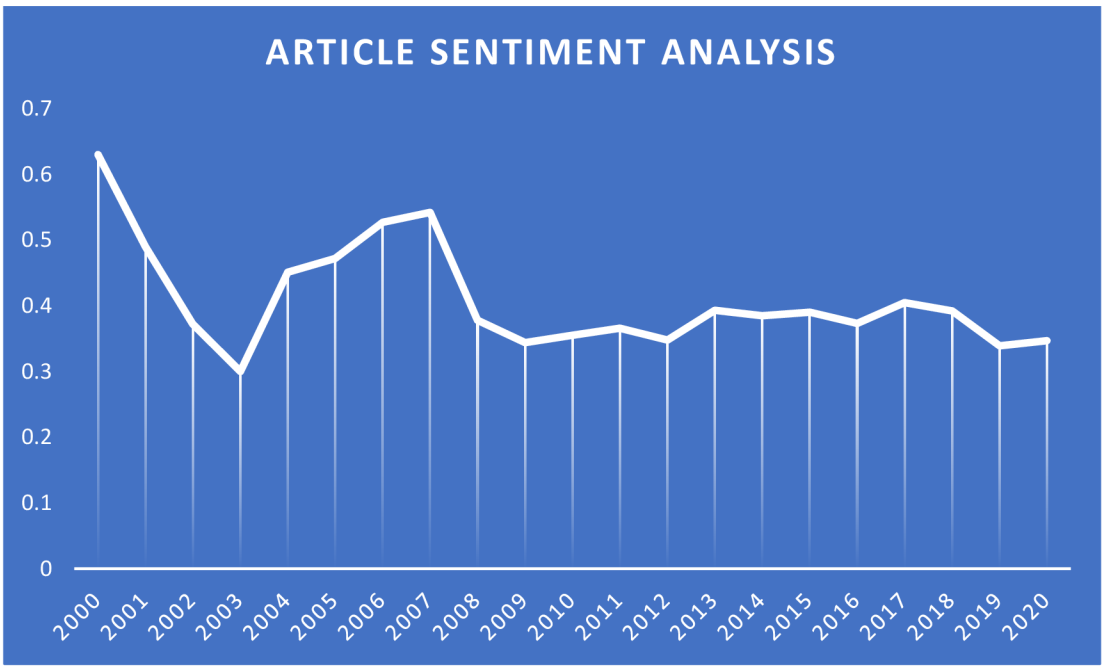

Figure 8: Article sentiment analysis (1 being positive, 0 neutral, -1 negative)

Results Interpretation

From the Figure 8 we can see the positive sentiment peaks in years 2000 and 2006 with 0.63 and 0.53 respectively. The sentiment is then the most negative in the year

2003 with 0.3 score. The sentiment is also relatively steady from 2008 onwards. It is possible that these negativity peaks were caused by Iraq war in 2003 and economic crisis in 2008, however this is only speculation considering the limited depth of our analysis and would likely require much deeper analysis to prove.

The difference between the first and final year of our analysis is rather severe with the score of 0.63 in 2000 and 0.35 in 2020, marking a shift of 0.28 points towards negativity. There is also a small difference between the two periods of 2000 – 2010 (0.44 avg) and 2011 – 2020 (0.37 avg) with a shift of 0.05 points towards negativity. The year 2016 does not seem to affect the sentiment in any way as the trend is rather steady after 2008.

Notably, while the VADER tool might not be the most reliable, it evaluates the articles on average with the score of 0.42, which is a rating between positive, and neutral, leaning slightly towards neutral.

Summary

Article sentiment analysis	
2020 vs 2000	***0.28 points shift towards negativity***
2000 – 2010 vs 2011 – 2020	***0.07 points shift towards negativity***
2000 – 2015 vs 2016 – 2020	***0.05 points shift towards negativity***
Accuracy	***Low***
Relevance	***High***
Trend of tabloidization	***Yes***

Table 8: Article sentiment analysis – summary

4.7 Q6: HOW DID THE DIFFERENCE IN THE SENTIMENT OF THE TITLE AND CONTENT CHANGE?

Pseudocode

```
# setup the VADER analysis tool
# train the VADER with the native provided corpora
```

```
# for each data entry:
# # if the article is purely audiovisual, skip the article from the analysis
# # if the article does not have a body or a title, skip the article from the analysis
# # get article body compound score with the VADER
# # get article title compound score with the VADER
# # subtract the title compound from the article compound to get the difference
# # get the difference average for the year
```

Relevance

While this indicator is not as frequently explored as others, we already established three factors that make the difference between the sentiment of the title and sentiment of the article **highly relevant**. Firstly, we recognized that increased negativity correlates with user engagement and draws to online articles much needed visitors. Secondly, we established that media are very often presented only as a combination of a thumbnail, title, and lead in the concurrent fast-pace world, especially on mobile devices. Thirdly, we introduced the notion of *clickbaits*, which exploit a completely different sentiment of the title (usually very negative) from the sentiment of the article (usually less negative) that became popular due to this logic. Analysing this parameter would allow us to get a glimpse of whether the WSJ is perhaps following an approach based on the same logic to attract readers.

Accuracy

While we assumed approximately 60% accuracy in the case of Q5, we must arguably assume even lower accuracy in this case. Since the parameter is obtained by subtracting the article title compound score from the article body compound score, the obtained number signals the difference in the sentiment. A positive result means the article is more positive than title. A negative number means the article is

less positive than the title[21]. Right from the start we are facing an issue that WSJ keeps the article titles extremely objective, and there is very little indication of any trend towards *clickbaits* or provocation, as titles such as *President Putin; Asia Briefs; 1999's Megamergers; Online Discussions*; are extremely common despite the presence of more expressive titles such as *E-Commerce Casts a Shadow Of a Doubt After Golden Run*. Because of this, VADER evaluates a big portion of the titles as simply neutral with compound score of 0.0 (which is arguably correct). Since we established in the Q5 that most of the articles are between neutral and positive, we can already conclude based on these two facts that a big portion of results will therefore yield a positive number – which in our case means that the article is more positive than the title (result of 0.5 means the article is 0.5 points more positive than the title), so while the goal is to look for any trends, we can already expect significant difference in the sentiments by default.

As such, we expect this analysis to be the least accurate out of all the research questions. Therefore, we assume this parameter to be of a **very low accuracy**.

[21] Example: on the scale where -1 is negative and 1 is positive, a positive article with 0.8 score with a negative title of -0.5 score will give us *(0.8 – (- 0.5))* = *1.3*, meaning the article is 1.3 points more positive than the title on the scale where

Run Chart

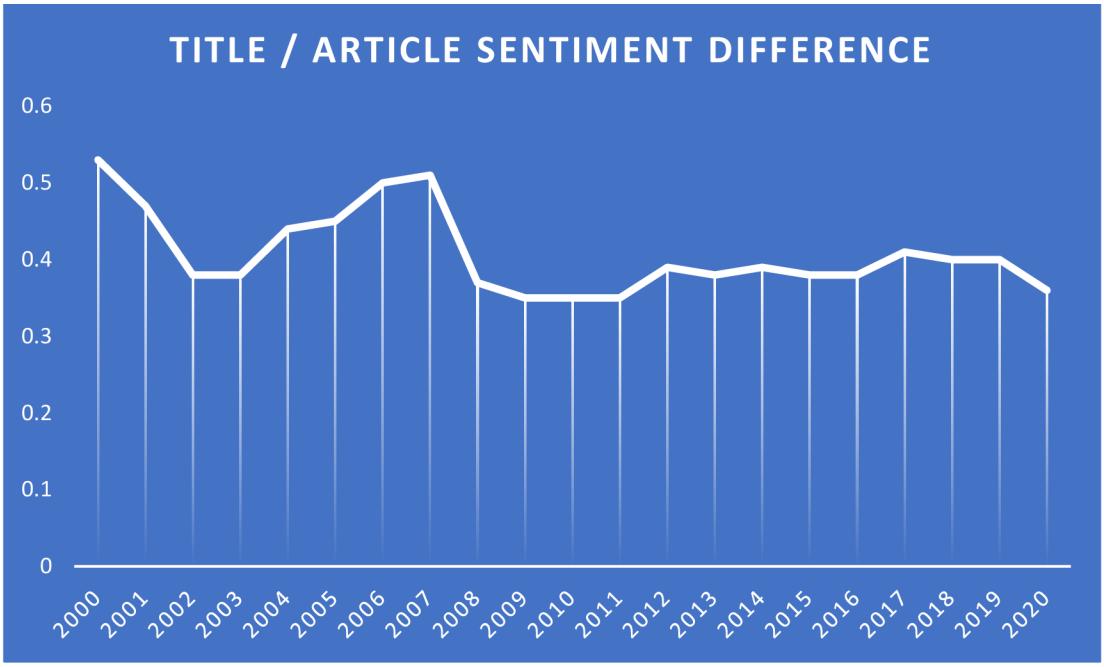

Figure 9: Difference in the article sentiment and title sentiment (value represents the difference utilizing VADER score)

Results Interpretation

From the Figure 9 we can see that the run chart is almost identical to the run chart from the Q5 with the two peak years when the difference between the two sentiments was the highest being 2000 (0.53) and 2006 (0.5) and steady trend after 2008. There is a difference between the year 2000 (0.53) and 2020 (0.36) marking a 32% decrease, meaning the difference between the sentiment of the title and the sentiment of the article body decreased over the years – the opposite of tabloidization. There is also decrease between the periods of 2000 – 2010 (0.43 avg) and 2011 – 2020 (0.38) marking 11.6% decrease.

Notably, when comparing the chart with the chart from the Q5, the average sentiment of the article being 0.42 and the average difference between the article body and article title sentiment being 0.4, we can determine that the article titles are almost purely neutral hovering around the sentiment of .02, which means the average title of a WSJ article is very objective and stripped of any sentiment whatsoever.

Summary

Article body and title sentiment difference	
2020 vs 2000	***0.17 points decrease***
2000 – 2010 vs 2011 – 2020	***0.05 points decrease***
2000 – 2015 vs 2016 – 2020	***0.02 points decrease***
Accuracy	***Very Low***
Relevance	***High***
Trend of tabloidization	***No***

Table 9: Difference in the article sentiment and title sentiment – summary

4.8 Q7: HOW DID THE COUNT OF PUNCTUATION SYMBOLS (?!) IN THE TITLE CHANGE?

Pseudocode

```
# for each data entry:
# # do not skip audiovisual articles
# # go through an article header and detect whether it contains "?" or "!"
# # calculate the average of articles containing the punctuation symbols per year
```

Relevancy

We have established that the title together with the thumbnail (cover photo) are frequently two of the main attention-capturing devices in today's digital world. However, WSJ does not really utilize frequent quotation or exclamation marks in their title and from the Q6 we know that the titles are relatively neutral throughout the years. Despite this, increased trend of present punctuation could be considered quite relevant indicator of tabloidization. Overall, we consider this indicator to have **medium relevancy** within our research,

Accuracy

Given the nature of this process – the fact we managed to collect all the titles, and the fact that to detect **?!** characters programmatically is very easy, we consider this analysis to be **very highly** (**100%**) **accurate**.

Run Chart

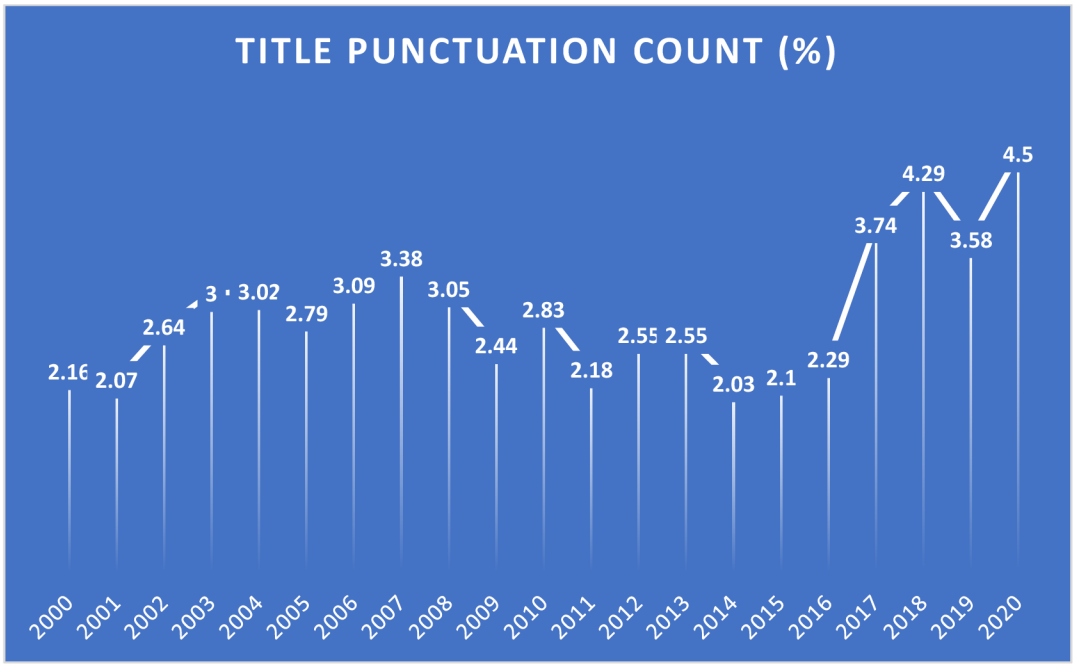

Figure 10: Title punctuation analysis (value represents the percentage of articles containing "!" or "?" symbol).

Results interpretation

From the Figure 10 we can see that on average only 2.87% of articles contain a question or quotation mark in the title, not surprising given the neutrality of the WSJ titles we confirmed in the Q6. While the punctuation percentage is somehow consistent (2.77 in the period of 2000 – 2010 and 2.98 in the period of 2011 – 2020, marking 7.6% increase), there is a quite distinctive trend of growing punctuation after the year 2016 (2.62% before the year 2016 and 3.68% in 2016 – 2020, marking 40.5% increase). The difference between the first year (2.16%) and last year, which is the peak year with 4.5%, is also quite significant with 108% increase.

Overall, we can detect sharp increase in the period 2016 – 2020, which might be caused by the switch in the company policy towards digital first approach.

Summary

Article body and title sentiment difference	
2020 vs 2000	***108% increase***
2000 – 2010 vs 2011 – 2020	***7.6% increase***
2000 – 2015 vs 2016 – 2020	***40.5% increase***
Accuracy	***Very High***
Relevance	***Medium***
Trend of tabloidization	***Yes***

Table 10: Article body and title sentiment difference – summary

4.9 Q8: HOW DID THE PERCENTAGE OF POLITICAL / ECONOMIC NEWS CHANGE?

Alternative solution

While our initial plan was to simply calculate the number of articles not belonging into *Politics* or *Economy* section, this eventually turned out to be not feasible as articles before 2010 were simply not assigned any section and even the articles after 2010 seem to have sections assigned inconsistently, leading to inconsistent results and irrelevant interpretation.

As a *workaround* we chose to implement a solution utilizing lexical markers. The presence of any of these lexical markers within the article will signal that the article concerns itself with political or economic context. Our goal then changes to measuring percentage of such articles. Note that these lexical markers are selected specifically for WSJ and consist of several terms that seem to be most frequently present in their respective categories. This solution is by no means optimal, it is very simplistic by its nature, and it is customized for the discourse of

WSJ. Nevertheless, after attempting this analysis with several permutations of these (or different) lexical markers, and reaching very similar results every time, we found this solution sufficient to serve as an acceptable alternative.

We selected the following lexical markers:

Politics: [‘politic’, ‘democrat’, ‘republican’, ‘president’, ‘govern’]

Economy: [‘econom’, ‘market’, ‘stock’, ‘business’, ‘profit’]

Pseudocode

```
# for each data entry:
# # skip audiovisual articles
# # merge article header, sub-header, description, and article full text
# # check whether the article contains any of the markers
# # calculate the average percentage of marked articles per year
```

Relevancy

We established the ratio of soft news and hard news as probably the most prominent indicator of the range dimension and probably one of the most important indicators of tabloidization overall. However, given the fact that we had to resort to the alternative solution as our primary solution is not available, we are going to treat this analysis as being of **low relevancy.**

Accuracy

As pointed out when describing the alternative solution, this method is rather a workaround serving as a compensation for corrupted dataset caused by archive inconsistency rather than exact indicator of the type of news. And while from programmatical point of view this analysis is 100% accurate, we will assume this analysis to be of **low accuracy.**

While accuracy of our analysis is arguably imprecise, it is again important to stress out that rather than to achieve a perfect process, our goal is to spot a trend, which should show up over a large dataset.

Run Chart

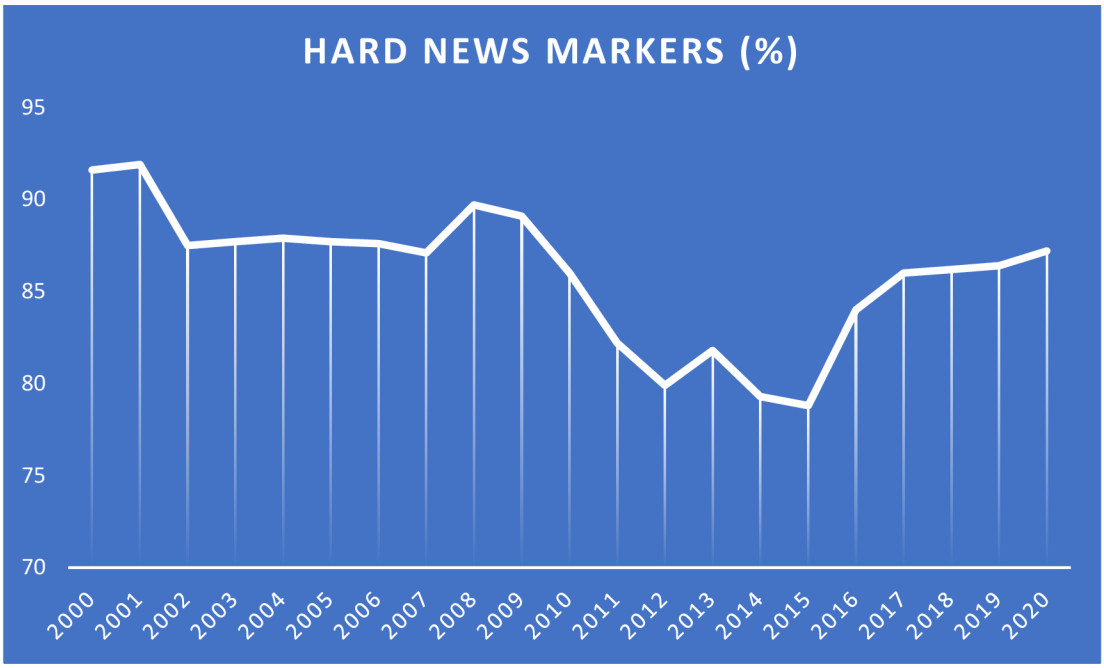

Figure 11 Hard News Markers (value represents the percentage of articles where at least one of the markers is present)

Result Interpretation

From the Figure 11 we can observe a rather small difference between the year 2000 (91.6%) and the year 2020 (87.2%) marking a mere 4.8% decrease. Probably the most notable part of the chart is the dip before the period of 2016 – 2020, which goes against our expectations of potentially increased trend of tabloidization after the switch to digital-first approach. Overall, there is basically no difference between the periods of 2000 – 2015 (85.99%) and 2016 – 2020 (85.96%). Likewise, there is relatively very small difference between the two periods of 2000 – 2010 (88.53%) and 2011 – 2020 (83.18%) with a small drop of 6%.

Overall, we observe just a very minor indication of tabloidization. In combination with arguably suboptimal work-around we chose to implement for this

analysis, we can effectively rule out any presence of tabloidization within this analysis.

Summary

Hard News Markers	
2020 vs 2000	***4.8% decrease***
2000 – 2010 vs 2011 – 2020	***0% change***
2000 – 2015 vs 2016 – 2020	***6% decrease***
Accuracy	***Very low***
Relevance	***Low***
Trend of tabloidization	***No***

Table 11: Percentage of hard news – summary

4.10 INDICATORS SUMMARY

Indicator	**Relevancy**	**Accuracy**	**Trend**
Content length	High	High	No
Images number	Low	High	No
Videos number	Low	High	No
AV space allocation	**High**	**High**	**Yes**
Sentiment	**High**	**Low**	**Yes**
Title / content sentiment difference	High	Very Low	No
Title punctuation (?!) count	**Medium**	**Very High**	**Yes**
Hard news ratio	Low	Low	No

Table 12: Analysed indicators - summary

5 DISCUSSION

In the following section we will attempt to describe some of the areas which could be improved upon when conducting a similar research as well as to conclude the results of our research.

5.1 RESEARCH LIMITATIONS

One of the limitations of our research is that our dataset does not include all the articles WSJ archive offers. Our initial goal was to scrape at least 10,000 articles per year. We eventually managed to scrape on average 36,042 articles per year with 756,889 articles in total – more than sufficient sample size. However, the dataset is not fully complete. Due to technical factors (e.g., faulty internet connection on both sides or unforeseen bugs restarting the crawler late in the process) we managed to get a complete dataset only in about half of the years we were analysing. In the remaining years we simply did our best to scrape as many articles as possible. We estimate we collected about 75% of the WSJ Archive. Note that the missing data should have very little impact on the results, as even the year with the smallest dataset (2019) still contains a respectable number of 18,237 articles – a sufficient sample size to produce relevant results. It is, however, likely that this lack of full dataset introduces small variance to the results.

Another potential improvement of this research would be to have a dataset of a tabloid to compare the values with. Such a comparison would allow us to clearly see the discrepancies between specific parameters and make the trends more comprehensible. While the original intention was to conduct this research in such a manner, we did not manage to find a suitable digital archive belonging to a tabloid. Majority of the large tabloid journals simply do not possess digital archive in a format that would be processible programmatically. The archive of *Daily Mail*[22] is the closes to the suitable candidate; however, the archive is almost non-existent before the year 2010 and incredibly inconsistent in the years soon after. While being easily processible, it is quite likely that *Mail's* a dataset would not produce any

[22] www.dailymail.co.uk/

relevant results until the mid-2010s. Despite this, we believe that such a dataset would have only a comparative value, and that it does not decrease the relevancy of our research.

Finally, the parameters we chose to analyze, and the methods of our analyses do have flaws on their own. While we did our best to try to conduct our analyses as objectively as neutrally possible, it is sometimes possible to reach different results by slightly modifying the process (e.g., by including the *article header* in the analysis of audiovisual material such as videos and images, the research is more likely to display a positive trend over the years). It is also possible to configure VADER sentiment analysis tool in a different way, leading to slightly different results. Finally, workaround implemented in the Q8 is likely suboptimal device for hard news detection and serves just as a substitute due to corrupted dataset. Although the relevancy and accuracy of individual parameters differs across the research, we did our best to treat them as objectively as possible.

As mentioned, the goal of this research was not necessarily to develop a flawless process of detecting indicators of tabloidization, but rather to display a trend, while the large size of the dataset compensates for potential flaws on the side of the quality of the process, which we believe was achieved given the enormous dataset and the yielded results.

5.2 CONCLUSION

This diploma thesis focused on analysis of indicators of tabloidization within the WSJ over the 21-year period of 2000 – 2020 via quantitative analysis. WSJ is an American broadsheet focusing mostly on news of economic and politic nature.

We chose WSJ predominantly because of a very consistent and well-maintained online archive as well as the shift in company values in 2016 when the company opted for *digital-first* approach and announced several changes in the structure of the news.

First, we set the theoretical framework in which we defined three dimensions of tabloidization based on the theoretical approach of MacLachlan and Golding, we then used this framework to enumerate various indicators utilized by various researchers when exploring the phenomenon of tabloidization. We then

defined our own criteria based on which we cherry-picked several indicators that we found suitable for computer analysis. We set 8 research questions in total with the goal of analyzing 8 parameters to spot potential trend of tabloidization over the analyzed period.

We chose quantitative analysis for processing the data. The dataset was gathered by a crawler with which we managed to web-scrape over 750,000 articles from the WSJ online archive. The idea behind the very large dataset was that because the WSJ adheres very strictly to the definition of broadsheet, any trend of tabloidization would be very subtle and would have to be detected over a very large sample size, as well as to mitigate some of the variance and inaccuracies caused by the analyzer utilizing programming tools of various accuracy and reliability (e.g., VADER sentiment analyzer).

Out of the initial 8 research questions, we managed to detect the trend indicating tabloidization only in 3 instances. We detected the **increase in area occupied by audiovisual material** by 248% (4.5% in 2000 and 15.68% in 2020), which represents quite significant increase in reliance on videos and images within the article content. Despite the amount of text remaining relatively constant throughout the years, video material and images are now occupying several times larger area of the article body than in 2000. We also detected **increase in negativity** / reduce of positivity of articles by 0.28 points (0.63 in 2000 and 0.35 in 2020) over the years, which might be indicating tabloidization, as negative articles generate more reader attention, but it may also be linked to short-term or long-term political and economic affairs development. Finally, we detected **increase in punctuation (!?) count in the title** by 108% (2.16% in 2000 and 4.5% in 2020). This is another potential trend of tabloidization, as these punctuation symbols are characteristically absent from broadsheet titles, and increase might signal the increased reliance on catchy and attention-grabbing title in the digital age.

While we detected several other trends, such as increase in image count or increase in the number of videos (especially in the most recent years), it is hard to view these trends as indicators of tabloidization, as the amount length of the text has not changed throughout the years and these occurrences might be a consequence of technological progress and better bandwidth capacity and smartphone capabilities over the years.

In other areas WSJ remained surprisingly consistent over the years, namely the **content text-length** remains very stable even after 2016, which marks the promise of WSJ to "cut down on unnecessary words". Together with the incredible neutrality of the article titles analyzed with the help of the VADER tool these are two of the parameters that speak volumes about the high standards of the journal.

Overall, we cannot conclusively claim that WSJ is undergoing the process of tabloidization over the analyzed period. There seems to be a **severe increase in audiovisual material** and **very minor signs of tabloidization** in other areas (article negativity and punctuation symbols in the title), which could be caused by the efforts of the journal to capture more of the readers' attention or increase readers' engagement on the webpage. Despite this, given the fact the length of text as well as several other areas remain incredibly consistent and/or neutral throughout the years, there seems to be consistently very high standard of news production in place that remains stable over the years.

Our findings seem to be in accordance with other studies. In general, authors exploring tabloidization have generally found only minor proof of tabloidization or managed to confirm only small portion of hypotheses they set to explore (Jelínková 2019, Zapletalík 2020, Uribe and Gunter 2004).

We cannot reliably confirm the process of tabloidization based on the results of our analysis. There are some areas where some practices of WSJ are indeed converging with the practices implemented by tabloids, however we cannot claim that these features are lowering the quality of articles in any way. If anything, it appears that WSJ is implementing some of the tabloid features to enhance the readers' experience, improve user engagement, and supply the information presented with new audiovisual material.

Finally, we believe that our research provides valuable insight due to exploration of linguistic issues with non-linguistic tools. We were particularly impressed by the VADER sentiment analyzer, utilizing AI and machine learning to analyze natural language. We therefore believe our research offers a valuable framework utilizing non-standard tools that will surely become more and more relevant in the upcoming years.

6 APPENDICES

6.1 CODE REPOSITORIES

Crawler: https://github.com/Han1s/WSJ_spider

Analyser: https://github.com/Han1s/WSJ_analyser

6.2 DATA ENTRY EXAMPLE


```
{
  "id": 12,
  "url": "https://www.wsj.com/articles/buttigieg-campaign-raised-24-7-million-in-fourth-quarter-11577884648",
  "header": "Buttigieg Campaign Raised $24.7 Million in Fourth Quarter",
  "subheader": "With about $76 million in total last year, the former South Bend, Ind., mayor was one of the best fundraisers in the Democratic presidential field",
  "description": "With about $76 million in total last year, the former South Bend, Ind., mayor was one of the best fundraisers in the Democratic presidential field",
  "article_text": "Pete Buttigieg raised $24.7 million in the final three months of the year, his campaign said Wednesday.That brings the former South Bend, Ind., mayor's total last year to about $76 million, making him one of the best fundraisers in the Democratic presidential field.Other candidates had not yet announced their fundraising hauls for the fourth quarter, which ended Tuesday night, but some had teased information about their finances. Campaigns have until the end of January to disclose their quarterly fundraising totals.Joe Biden said as December began he had already hit $15 million for the quarter, about what he raised in the entire three-month summer fundraising period. In a more recent fundraising email, his campaign said the former vice president hoped to top his second-quarter total of about $22 million.Elizabeth Warren's campaign said in a fundraising email last week that the Massachusetts senator had raised $17 million for the quarter, a slowed pace from her third quarter, when she brought in $24.6 million.Tech entrepreneur Andrew Yang said on social media that his campaign collected more than $4 million from online donors just in the last few days.Mr. Buttigieg's campaign said more than 733,000 people have made a total of two million donations. The average contribution amount was $38, the campaign said.Mr. Buttigieg has aggressively pursued donors who can give the legal maximum of $2,800, a fundraising style like Mr. Biden's and most previous major presidential candidates. Two other top contenders this time, Vermont Sen. Bernie Sanders and Ms. Warren, have sworn off fundraisers with wealthy donors.Ms. Warren at a Los Angeles debate last month criticized Mr. Buttigieg for his courting of wealthy donors, drawing a response from her rival that she raised money in the same way for almost all of her political career. Write to Julie Bykowicz at julie.bykowicz@wsj.com",
  "section": "Politics",
  "type": "article",
  "article_type": "POLITICS",
  "keywords": "Political/General News,Politics/International Relations,Domestic Politics,Regional Politics,SYND,WSJ-PRO-WSJ.com,Politics & Policy,Pete Buttigieg,political,general news,politics,international relations,domestic politics,regional politics,politics & policy",
  "word_count": "315",
  "image_address": "https://images.wsj.net/im-140653/social",
  "image_count": "1",
  "image_width": "1280", "image_height": "640",
  "header_image_size": {"height": 413, "width": 620},
  "content_size": {"height": 1937, "width": 620},
  "article_content_image_sizes": [],
  "article_content_image_count": 0,
  "video_count": 1,
  "is_audiovisual": false
}
```


Figure 12: An example of a data entry

7 SHRNUTÍ

Cílem této práce je kvantitativní strojová analýza indikátoru bulvarizace na vzorku 750 000 článků pocházejících z deníků Wall Street Journal (WSJ). Za účelem této analýzy jsme využili programátorské nástroje a knihovny založené na programovacím jazyce *Python.* Tento proces byl rozdělen do dvou částí. V první části jsme vytvořili program určený ke sběru dat (*crawler*), jehož účelem bylo projít digitální archiv WSJ a sesbírat data ve vhodné podobě. Pro druhou část jsme vytvořili analyzační program (*analyzér*), jehož cílem bylo identifikovat a změřit předem vytyčené parametry za účelem identifikace trendů bulvarizace. Je zde zmíněno, že cílem práce není dosáhnout perfektní identifikace bulvarizace za využití strojové analýzy, ale spíše detekce trendů i za cenu méně přesného charakteru analýzy, který by měl být vykompenzován velikostí vzorku dat.

Úvodní teoretická část si klade za cíl vymezit základní koncepty, jako je *bulvár, seriózní tisk,* a *bulvarizace*, které slouží jakožto základ pro část metodologickou. Tato část se také zabývá vývojem bulvarizace, důležitostí kulturního a historického kontextu při zkoumání bulvarizace, a faktory způsobující bulvarizaci. Potenciální negativní či pozitivní vliv bulvarizace je pak diskutován na konci kapitoly.

Metodologická část se opírá o tři dimenze bulvarizace (*formu*, *styl* a *obsah*), ve kterých je možné bulvarizaci zkoumat. V rámci těchto tří dimenzí jsou zde pak vyčteny různé indikátory bulvarizace, které zkoumali jiní autoři zabývající se podobnou problematikou. Následně jsou zde stanovena tři kritéria (*relevance*, *přesnost* a *proveditelnost*) na základě kterých jsme z vyčtených indikátorů vybrali indikátory vhodné pro strojovou analýzu bulvarizace. Tyto indikátory pak daly za vznik osmi výzkumným otázkám zkoumajících osm parametrů za účelem detekce trendů bulvarizace:

1. Zkrácení délky textu
2. Nárůst počtu obrázků ve článku
3. Nárůst počtu videí ve článku
4. Nárůst podílu audiovizuálního materiálu v poměru k textu článku
5. Nárůst negativního sentimentu článku
6. Nárůst negativního sentimentu titulku (v porovnání se sentimentem článku)
7. Nárůst podílu článku obsahujících v nadpise symboly ? a !

8. Pokles podílu ekonomických a politických zpráv

V této části práce je také stručně popsán charakter deníku WSJ a změna v oblasti hodnot deníku za účelem digitalizace deníku v roce 2016. Je zde také popsán výzkum podobného charakteru jak v oblasti bulvarizace, tak v oblasti strojové analýzy jazyka. Konec sekce pak představuje technologii, která byla v této práci využita“

- programovací jazyk *Python* určený k datové analýze a práci s umělou inteligencí
- knihovnu učenou ke sběru dat z internetu (tzv. *web-scraping*) *Selenium*,
- softwarový balíček určený ke zpracování přirozeného jazyka *Natural Language Toolkit*
- analyzér sentimentu *VADER* patřící k *NLTK*, postavený na bázi umělé inteligence a strojového učení

Následující kapitola se věnuje samotné strojové analýze. Nejprve je zde představen proces sběru dat využívající *web-scraping* v podobě programu využívající knihovnu *Selenium.* Tento program se snaží v optimalizované podobě získat co největší vzorek dat z digitálního archivu WSJ.

Ve druhé polovině kapitoly je pak daný vzorek dat analyzován *analyzérem* psaným v jazyce *Python* za účelem detekce indikátorů bulvarizace v rámci stanovených výzkumných otázek. U každé výzkumné otázky jsou podrobněji rozebrány následující části:

1. *Pseudokód* – tato část si dává za úkol rozebrat funkcionalitu analyzéru ve stručné a jasné podobě, snadno uchopitelné bez jakýchkoli technických znalostí na straně čtenáře práce.
2. *Relevance* – tato část diskutuje relevanci daného indikátoru v kontextu deníku WSJ a výzkumu bulvarizace jako takového.
3. *Přesnost* – tato část diskutuje předpokládanou přesnost dané analýzy, případné nedostatky, a případný prostor pro zlepšení.
4. *Bodový graf* – tato část představuje výsledky dané analýzy ve formě přehledného bodového grafu.

5. *Interpretace výsledků* – tato část rozebírá výsledná data, snaží se na jejich základě stanovit, zda dochází k bulvarizaci a zkoumá možné příčiny a nedostatky dané analýzy.
6. *Shrnutí* – tato část nabízí stručný souhrn dané analýzy ve formě tabulky.
7. *(Představení nástroje – pouze u analýzy využívající VADER)* – tato část představuje analyzér sentimentu VADER, stručně ukazuje jeho využití a diskutuje jeho přesnost a použitelnost při strojové analýze textu.
8. *(Alternativní řešení – pouze u otázku č.8)* – tato část nabízí alternativní řešení pro otázku č.8, které bylo nutné implementovat kvůli nevalidnímu vzorku dat na straně WSJ.

Z daných osmi otázek se trend bulvarizace potvrdil pouze u třech. U *nárůstu podílu audiovizuálního materiálu v poměru k textu článku*, kde došlo ke 248% (4.5% v roce 2000 a 15.68% v roce 2020) navýšení při porovnání prvního a posledního roku datového vzorku, *nárůstu negativního sentimentu článku*, kde došlo ke změně o hodnotě 0.28 bodu (0.63 v roce 2000 a 0.35 v roce 2020) a *nárůstu podílu článku obsahujících v nadpise symboly ? a !*, kde došlo ke 108% nárůstu (2.16% v roce 2000 a 4.5% v roce 2020). Překvapivě nebyl potvrzen trend *zkrácení délky textu,* který byl předpokládán na základě změny v hodnotách WSJ v roce 2016. Analýza vedla ale i k zjištěním v oblasti nárůstu počtu videí a obrázků, což jsme ale v daném kontextu nepovažovali za indikátor bulvarizace, jelikož se nezměnila délka textu, a překvapivě téměř dokonalé neutralitě titulků článků.

V následující kapitole jsou pak rozebrány možné nedostatky našeho výzkumu, zejména částečně nekompletní vzorek dat, potenciální nepřesnosti v analyzéru a suboptimální alternativní řešení otázky č.8. Jsou zde také shrnuty a interpretovány výsledky analýzy. Na základě získaných výsledků není možné tvrdit, že by WSJ podléhal vlivu bulvarizace, a to i navzdory radikální změně ve firemní politice v roce 2016. Výsledná data poukazují na velmi vysoký standard WSJ, neutralitu i navzdory vlivům moderních technologií a změnám ve spotřebitelském chování. V deníku dochází ke většímu důrazu na audiovizuální složku, jakou je počet videí a obrázků, což ale může být zapříčiněno technologickým vývojem, zvyšující se dostupností a kvalitou internetového připojení a možnostmi mobilních telefonů. Na základě získaných dat byla tedy

bulvarizace detekována jen ve velmi malé míře a nelze hovořit o poklesu standardu deníku.

Věříme, že daný výzkum umožnil poněkud odlišný úhel pohledu na zkoumání lingvistických otázek za využití moderních nástrojů. Překvapivým prvkem byl zejména analyzér sentimentu VADER, který využívá umělé inteligence a strojového učení k analýze přirozeného jazyka. Věříme, že hodnota našeho výzkumu spočívá i ve využití těchto nestandardních nástrojů, které se do budoucna budou zajisté stávat stále relevantnějšími.

8 BIBLIOGRAPHY

8.1 REFERENCES

Ahrens, Joseph. "The decline in newspapers: A closer look." *Wake Review Literary Magazine & Club, Wake Tech Community College*, November 2016.

Baig, Edward. "Too scary? Elon Musk's OpenAI company won't release tech that can generate fake news." *eu.usatoday.com.* 2 15, 2019. https://eu.usatoday.com/story/tech/2019/02/15/elon-musks-openai-fake-news-generator-too-scary-release/2880790002/ (accessed 11 11, 2021).

Bek, Mine Gencel. "Research Note: Tabloidization of News Media: An Analysis of Television News in Turkey." *European Journal of Communication* 19, no. 3 (2004): 371-386.

Bennett, W. Lance. *News: The politics of illusion.* University of Chicago Press, 2016.

Beri, Aditya. „Sentiment Analysis Using Vader.“ *towardsdatascience.com.* 27. 5 2020. https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664#:~:text=VADER%20(%20Valence%20Aware%20Dictionary%20for,directly%20to%20unlabeled%20text%20data.).

Bird, Elizabeth. "News We Can Use: An Audience Perspective on the." *Javnost - The Public*, 1998: 33-49.

Bird, S. Elizabeth. "Tabloidization: What is it, and does it really matter?" (Routledge) 2009: 40-50.

Bowden, John. *Wall Street Journal editorial: Conservatives 'could live to regret' Trump emergency declaration.* 11. 01 2019. https://thehill.com/homenews/media/424877-wall-street-journal-editorial-conservatives-could-live-to-regret-trump (přístup získán 5. 10 2021).

Brandelid, Annie, and Evelina Eklund. *Tabloidization in Swedish news media? The ongoing pandemic in focus.* Thesis, Karlstads University, 2021.

Connell, Ian. "Mistaken identities: Tabloid and broadsheet news discourse." *Javnost-the public*, 1998: 11-31.

Esser, Frank. "`Tabloidization' of News: A Comparative Analysis of Anglo-American and German Press Journalism." *European Journal of Communication* 14, no. 3 (1999): 291-324.

Fang, Irving. *A history of mass communication: Six information revolutions.* Routledge, 1997.

Heitzmann, Philippe. "Web-scraping Wall Street Journal articles for sentiment analysis." *nycdatascience.com.* 09 21, 2020. https://nycdatascience.com/blog/student-works/scraping-wall-street-journal-article-data-to-measure-online-reader-engagement-an-nlp-analysis/ (accessed 08 1, 2021).

Hughes, Gary. *A History of the Broadsheet Newspaper.* 2 12, 2021. https://www.historic-newspapers.co.uk/blog/broadsheet-history/ (accessed 10 1, 2021).

Jelínková, Veronika. "Tabloidization of dailies Mladá fronta Dnes and Blesk in 2005-2017." Master Thesis, 2019.

Karlsson, Michael Bo. "Goodbye politics, hello lifestyle" Changing news topics in tabloid, quality and local newspaper websites in the U.K. and Sweden from 2002 to 2012." *Observatorio Journal* 10, no. 4 (2016): 150-165.

Košňárová, Barbora. *Language of Evaluation in Broadsheets and Tabloids.* Master Thesis, Palacký University Olomouc, 2018.

Kuhlman, Dave. *A Python Book: Beginning Python, Advanced Python, and Python Exercises.* 2012.

Lehman-Wilzig, Sam N, and Michal Seletzky. "Hard news, soft news,'general'news: The necessity and utility of an intermediate classification." *Journalism*, 2010, 11 ed.: 37-56.

Magin, Melanie, et al. "Is Facebook driving tabloidization." In *Global Tabloid: Culture and Technology*, by Martin Conboy and Scott A. Eldridge, 56-75. Routledge, 2021.

McLachlan, Shelley, and Peter Golding. "Tabloidization in the British press: a quantitative investigation into changes within British newspapers from 1952-1997." *Tabloid tales: global debates over media standards* (Hampton Press), 2000: 75-90.

Mullin, Benjamin. *poynter.org.* 11. 10 2016. https://www.poynter.org/business-work/2016/the-wall-street-journal-is-reorganizing-its-newsroom-and-cutting-down-on-flabby-stories/.

Myers, Margaret. "Qualitative research and the generalizability question: Standing firm with Proteus." *The qualitative report*, 2000, 4 ed.

Norris, Pippa. ""To Entertain, Inform, and Educate": Still the Role of Public Television." *Political Communication*, April 2001: 1-18.

Ochieng, Pamela A. "An analysis of the strengths and limitation of qualitative and quantitative research paradigms." *Problems of Education in the 21st Century 13* 13 (2009).

Otto, Lukas, Isabella Glogger, and Mak Boukes. "The Softening of Journalistic Political Communication: A Comprehensive Framework Model of Sensationalism, Soft News, Infotainment, and Tabloidization." *Communication Theoy* (International Communication Association) 27 (2017): 136-155.

Potthast, Martin, Sebastian Köpsel, Benno Stein, and Matthias Hagen. "Clickbait detection." *European Conference on Information Retrieval* (Springer, Cham), 2016: 810-817.

Rowe, David. "On going tabloid: A preliminary analysis." *Metro Magazine: Media & Education Magazine* 121/122 (2000): 78-85.

Rowe, David. "Tabloidization of news." 394-405. Routledge, 2009.

Rungta, Krishna. „What is Selenium? Introduction to Selenium Automation Testing.“ *guru99.com.* 7. 10 2021. https://www.guru99.com/introduction-to-selenium.html.

Salerno, Alison. "Is This Headline Clickbait?" *towardsdatascience.com.* 8 24, 2020. https://towardsdatascience.com/is-this-headline-clickbait-86d27dc9b389 (accessed 9 18, 2021).

Scanlan, Chip. *Writing from the Top Down: Pros and Cons of the Inverted Pyramid.* 2008. www.poynter.org (přístup získán 10. 10 2021).

Sellers, Frances Stead. "Embracing change: British dailies are trying a variety of new approaches in an effort to survive and thrive in a new media landscape. Are there lessons here for U.S. papers?" *American Journalism Review* (Gale Academic OneFile) 28, no. 5 (Oct-Nov 2006): 54+.

Smialer, Jeanna. "time.com." *Twitter Bots May Have Boosted Donald Trump's Votes by 3.23%, Researchers Say.* 5 21, 2018. https://time.com/5286013/twitter-bots-donald-trump-votes/ (accessed 11 11, 2021).

Spillane, Brendan, a et al. „Tabloidization versus credibility: Short term gain for long term pain.“ *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

Steiner, Miriam. "Soft Presentation of Hard News? A Content Analysis of Political Facebook Posts." *Media and Communication*, 2020, 8 ed.: 244-257.

Sterne, Peter. *Wall Street Journal goes digital-first, creates dedicated print desk.* 11 10, 2016. https://www.politico.com/media/story/2016/10/wall-street-journal-goes-digital-first-creates-dedicated-print-desk-004803/ (accessed 11 1, 2021).

Thussu, Daya Kishan. *News as entertainment: The rise of global infotainment.* Sage, 2008.

Uribe, Rodrigo, and Barrie Gunter. "Research Note: The Tabloidization of British Tabloids." *European Journal of Communication* 19, no. 3 (2004): 387-402.

Zapletalík, Radim. *The Influence of Tabloidization on Sport News in the Czech Republic and the USA*. Master Thesis, Olomouc: Palacký University, Olomouc, 2020.

8.2 ONLINE DICTIONARIES

Collins Dictionary, Accessed on December 5, 2021 https://www.collinsdictionary.com/dictionary/english/broadsheet

8.3 WEBSITES

Bbc.co.uk, Accessed on November 3, 2021. www.bbc.co.uk/bitesize/guides/zps4qty/revision/1

Chromedriver, Accessed on October 3, 2021. https://chromedriver.chromium.org/

Cision, Accessed on October 3, 2021. https://www.cision.com/2019/01/top-ten-us-daily-newspapers/¨

Daily Mail, Accessed on October 12, 2021. www.dailymail.co.uk/

Github, Accessed on December 2, 2021. https://github.com/cjhutto/vaderSentiment

Guardian, Accessed on October 2, 2021. https://www.theguardian.com/media/2015/jun/11/wall-street-journal-european-and-asian-editions-broadsheet

NTLK, Accessed on September 1, 2021. https://www.nltk.org/

Scrapy, Accessed on August 22, 2021. https://doc.scrapy.org/en/latest/intro/overview.html

W3Schools, Accessed on August 21, 2021. www.w3schools.com

Wall Street Journal, Accessed on December 10, 2021. https://www.wsj.com

9 LIST OF TABLES

10 LIST OF FIGURES

11 ABSTRACT

This thesis represents a quantitative research focused on indicators of tabloidization within Wall Street Journal during the period of 2000 – 2020 based on the dataset of more than 750,000 articles. The objective of the thesis was to programmatically collect the data in as optimized way as possible and subsequently process them with an analyzer written in python processing various parameters looking for indicators of tabloidization. The results are then presented in a form of a run chart and interpreted.

Keywords

Computer analysis

Wall Street Journal

Tabloidization

Indicators of tabloidization

Web scraping

Sentiment analysis

12 ANOTACE

Cílem této práce je kvantitativní výzkum zaměřen na indikátory bulvarizace v deníku Wall Street Journal během období 2000–2020 na základě vzorku překračujícího 750 000 článků. Práce využívá programovacího jazyka Python ke sběru dat v optimalizované podobě a následně využívá druhý analyzační program napsaný v totožném jazyce ke zpracování různých parametrů indikující bulvarizaci. Výsledek je prezentován ve formě bodového grafu a interpretován.

Klíčová slova

Strojová analýza

Wall Street Journal

Bulvarizace

Indikátory bulvarizace

Web scraping

Analýza sentimentu