FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Václav Stibor

Financial News Sentiment Analysis

Department of Software Engineering

Supervisor of the bachelor thesis: Supername Supersurname
Study programme: study programme
Study branch: study branch

Prague YEAR

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

Dedication. It is nice to say thanks to supervisors, friends, family, book authors and food providers.

Title: Financial News Sentiment Analysis

Author: Václav Stibor

Department: Department of Software Engineering

Supervisor: Supername Supersurname, department

Abstract: One key skill required to make good investments in the stock market is being able to correctly analyze news related to the finance and the business sector. Which company is diversifying its sectors or which company is showing signs of heading towards bankruptcy? You need to keep yourself updated with every little deal and fallout happening in the market. Financial news can be a little tricky to understand especially for those who are new to the financial world.

Keywords: key words

Contents

Introduction

In today's era of information explosion and constant flow of information, it becomes more time-consuming to keep track of associations and deeply understand the published content through media and online news, primarily when investing in a specific area. For instance, the investment in a company like Apple Inc. requires acquiring and processing a wide range of available information with significant effort and dedication in studying articles and other sources. At the same time, publicly available information resources such as news articles and tools like sentiment analysis allow us to transfer real-world context into the digital environment and use it for our benefit.

Sentiment analysis, the ability to identify and evaluate the emotional charge of content, has evolved into a crucial instrument for comprehending opinions, attitudes, and the general atmosphere surrounding various topics. This work focuses on developing an application that allows users to visualize connections between companies and news articles using a knowledge graph network and the impact of news sentiment on a company's stock price, even in real time.

Many experiments are currently being conducted based on historical data to examine the effect of sentiment, but not on current data, despite the rather promising results on datasets. The absence of such an application motivates this thesis. An application that extracts actual data from news articles for sentiment analysis and subsequently evaluates the future impact of that sentiment on a company's stock price.

This thesis will discuss the technical aspects of sentiment analysis and implementing an application that conveys this information to users as recently as possible. The aim is to provide users with a tool that allows them to actively monitor and analyze the flow of information about emotional overtones as one of the key identifiers in trading decisions. The thesis will be structured as follows. Chapter 1 will discuss the theoretical background behind the stock market. Chapter 2 will give an overview of data sources and the data itself. Chapter 3 will discuss the sentiment analysis and design of the application. Chapter 4 will discuss the implementation of the application. Chapter 5 will discuss the evaluation of the application. Chapter 6 will discuss the conclusion and future work.

1. Theoretical Background

Since the application's core is sentiment analysis, it is necessary to define the basic concepts. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus.

1.1 Sentiment Analysis Basics

Sentiment analysis or opinion mining is a subfield of Natural Language Processing (NLP) that aims to identify and extract opinions and emotions from a text. The goal is to determine the author's attitude towards a particular topic or the overall contextual polarity of various document levels. We measure the text's polarity using a numerical scale ranging from -1 to 1. The low-end score of the scale signifies a negative sentiment, zero represents neutrality, and the high-end score indicates a positive sentiment. This scale effectively estimates the degree of negativity or positivity in the text's tone.

The extraction of opinions and emotions has applications in various areas, from product reviews to political events. Hence, it is imperative to work in different domains (see Piryani et al., 2017). Because of cross-domain and cross-language, two of the most general issues in sentiment analysis, this thesis will focus only on the financial domain in English. Nevertheless, domain-specific sentiment analysis achieves remarkable accuracy while staying highly domain-sensitive, as shown in (Saunders, 2020). To delve deeper into cross issues, Liu provides further details in his book (Liu, 2022).

1.1.1 Levels of Sentiment Analysis

Sentiment analysis has been studied at several levels of granularity: Document-level, Sentence-level, Phrase-level, and Entity-level[1], as illustrated in Figure 1.1.

[1]Entities are sometimes referred to as targets, hence Target-level or Target-based sentiment analysis.

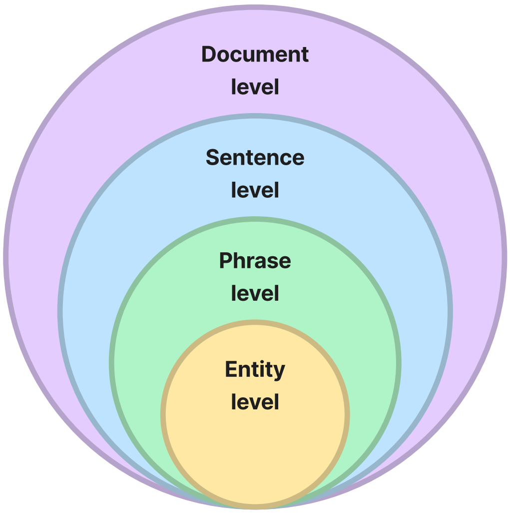

Figure 1.1 Levels of sentiment analysis (inspired by Wankhade et al., 2022).

Document-level

Document-level sentiment analysis is the most straight level. The task is to determine the overall emotional context of the entire document, such as a chapter, article, or review, whether or not involving a study of entities or aspects. This level gives us a general assessment of whether the content is more likely to be positive, negative, or neutral.

Sentence-level

Sentiment analysis at the sentence level focuses on individual sentences within the text. We observe the polarity of each sentence autonomously, employing the same methodologies utilized at the document level but with an increased volume of training data and enhanced processing resources. This level is more challenging than the document level because it requires a more in-depth understanding of the text.

Phrase-level

Phrase-level sentiment analysis examines sentiment within smaller linguistic units such as phrases or sentence members. Thus, it can better reveal the emotional charge in specific parts of sentences. Additionally, this level is more challenging than the sentence level because it requires a more detailed understanding of the text.

Entity-level

The most elaborative level is entity-level sentiment analysis, where we study sentiment associated with specific entities mentioned in the text. This level provides a detailed look at the expressed polarity of certain products, individuals, or organisations. One of the main tasks in this scope is the named entity recognition, which will be discussed later.

Some researchers classify the last level as the aspect-level, as noted by Wankhade et al., 2022, or a more detailed entity-level version called the feature-level proposed by Mary et al., 2017. While both approaches aim to evaluate sentiment towards specific aspects, they differ in their task approach. Relationships between these levels are illustrated in Figure 1.2.

In the first case, aspects are considered without directly mentioning entities in the text. We are not interested in the entities since the input textual data are commonly associated with them[2], such as reviews. The study conducted by Wang et al., 2019 analyzed sentiment at the aspect level within restaurant reviews. It primarily examines aspects such as food, price, service, and others. In the feature-based approach, aspects are commonly associated with an entity's features by connecting the entity and its aspects in text. To illustrate, consider the sentence:

"The battery life of this phone is excellent, but the camera is not good."

At the feature level, we identify *the battery life* and *camera* as specific features of entity *the phone*, allowing us to determine the polarity of each entity's feature.

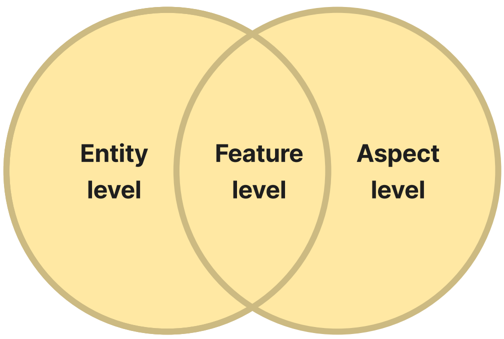

Figure 1.2 Comprehensive overview of the last level.

The term entity-level sentiment analysis is frequently employed in literature, and some studies consider it synonymous with targeted sentiment analysis,

[2]Entities are not handled in this case, but we provide them here for a better understanding.

as discussed Rønningstad et al., 2022 in the terminology review. For our purposes, entity-level sentiment analysis better captures the aggregate, document-wide approach, where a single entity can be associated with multiple targets in different sentences, discerning it from traditional target-level sentiment analysis.

However, this thesis primarily focuses on entity-level sentiment analysis, excluding consideration of the entity's features. This decision is motivated by treating the mentioned companies in news articles as entities rather than delving into their specific aspects. Additionally, entity and aspect extraction as separate tasks are complex and challenging, given that the methods and facets employed for recognition differ due to their distinct characteristics (Liu, 2015; Zhang et al., 2014).

TODO: Pokud nenarazím na článek, který by to vyvrátil. Navíc se zkoumáním aspektů by přibylo spousty práce.

1.2 Named Entity Recognition

Named entity recognition, or entity extraction, constitutes a fundamental component in NLP dedicated to identifying and classifying proper nouns into predefined semantic classes. These classes are unlimited since entities could be anything we can categorize using a tag, including the names of organisations, people, places, or other available information from unstructured textual data such as time, quantity, and money expressions. Someone well-versed in data analysis must have contemplated the possibility that some sources provide data in which the names of organisations are directly associated with their unique tags in text, but this is not our case (for more details, see Chapter 3). Understanding the importance of this task, we recognise its essential role in entity-level sentiment analysis, as it allows us to identify the emotion associated with specific entities mentioned in the text.

TODO: Vrátit se až budu mít všechny 3 zdroje a případně pozměnit na "né vždy náš případ."

Tim Cook **PERSON** was named the new CEO of Apple Inc. **ORG** on August 24, 2011 **DATE** .

Figure 1.3 TODO: Named entities along with their associated label classes.

The example above illustrates that a single entity can contain more than one word. This challenge is addressed by token tagging formats outlined in the paper (Ramshaw et al., 1995). Individual words are referred to as tokens. The formats describe the position of each token in a named entity, such as the Beginning-Inside-Outside (BIO) format, also known as the Inside-Outside-Beginning (IOB) format as well as other derived names like Inside-Outside (IO) and Beginning-Middle-End-Whole-Outside (BMEWO). These formats discharge us from the constraints of entities, which are hard-coded or unreliably specified by regular expressions.

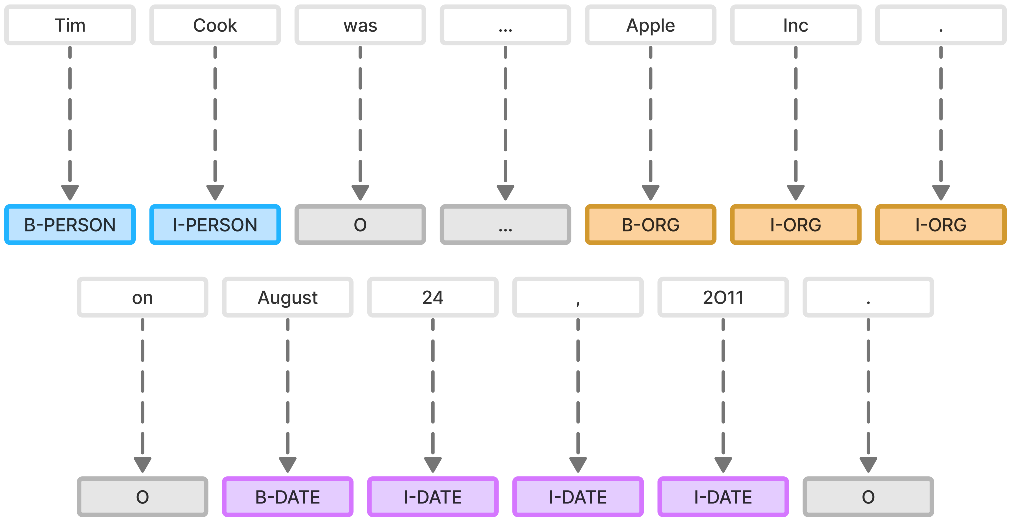

Figure 1.4 TODO - BIO2 format.

The tags that assign tokens to each entity class contain prefixes. The prefix I- indicates that the token is contained in the named entity, whereas O- indicates that the token is not contained in any named entity. The prefix B- is slightly more specific and indicates that the token is contained at the beginning of the named entity, followed immediately by a token not containing the O- prefix. A particular case of the BIO approach is the BIO2 format, denoting all tokens beginning with the prefix B-, regardless of whether a token with an O- prefix follows. A slightly more detailed approach is BMEWO, where the prefixes B- and O-, as in the previous ones, indicate the beginning and absence of the named entity occurrence, respectively. M- prefix symbolises the middle token between B- and E-, where the token with the prefix E- ends the named entity. Furthermore, the prefix W- indicates a single-token named entity. Table 1.1 below demonstrates how the mentioned sentence could be labeled with IO, BIO, BIO2, and BMEWO formats.

1.2.1 Techniques

TODO: Posunout k obrázku.

Approaches for entity extraction from unstructured data encompass a broad range of techniques, though they commonly converge into three principal categories. Specifically, the work (Keraghel et al., 2024) classifies them in the following way: Knowledge-based, Feature engineering, and Deep learning, as illustrated in Figure 1.5 below.

	Format			
Token	IO	BIO	BIO2	BMEWO
Tim	I-PERSON	I-PERSON	B-PERSON	B-PERSON
Cook	I-PERSON	I-PERSON	I-PERSON	E-PERSON
was	O	O	O	O
...	O	O	O	O
Apple	I-ORG	I-ORG	B-ORG	B-ORG
Inc	I-ORG	I-ORG	I-ORG	M-ORG
.	I-ORG	I-ORG	I-ORG	E-ORG
on	O	O	O	O
August	I-DATE	I-DATE	B-DATE	B-DATE
24	I-DATE	I-DATE	I-DATE	M-DATE
,	I-DATE	I-DATE	I-DATE	M-DATE
2011	I-DATE	I-DATE	I-DATE	E-DATE
.	O	O	O	O

Table 1.1 TODO

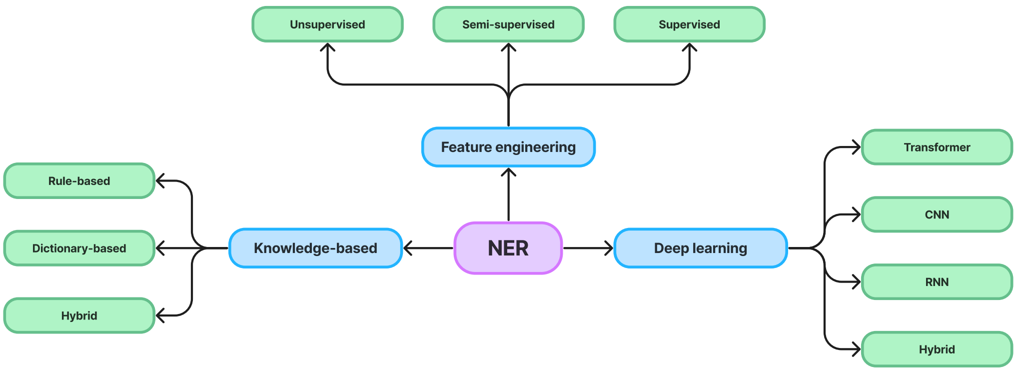

Figure 1.5 TODO: Named entity recognition main method approaches.

Knowledge-based

Knowledge-based approaches rely on predefined rules and dictionaries to identify entities. These rules and dictionaries, typically created by domain experts, recognise entities based on their characteristics. Rules can be given as regular expressions, denoting patterns to match character combinations in strings. Established on the chosen method, we separate this category into Rule-based and Dictionary-based or combine these approaches. The main advantage of knowledge-based approaches stems from their interpretability, as the rules and dictionaries can be easily understood and modified. Nevertheless, the principal disadvantage is frequent limitations to the entities in the manually created dictionaries and rules, as a result of which new entities cannot be recognised.

Feature engineering

Instead of manually creating a set of rules and a dictionary, feature engineering-based approaches, popularly identified as machine learning, use linguistic and statistical features to identify entities. These features are generally derived from the text and subsequently are used to train a machine learning model to recognise entities. The primary advantage of this approach is the ability to learn more about the data and discover patterns that may not be apparent at first. Additionally, their ability to recognise new entities differs from dictionaries in some cases. However, the main disadvantage is that these approaches require a large amount of labeled training data to be accurate.

Before exploring feature engineering-based methods in depth, we should clarify the data under discussion and the definition of labels. In the early introduction regarding named entity recognition, we referred to tags. These tags correspond to labels that serve as identifiers describing particular data that are consequently categorised according to the assigned label. To promote better comprehension, referring to the example illustrated in Figure 1.3, we have textual data in which labels are assigned as outlined in the following Table 1.2.

Data	Label
Tim Cook	**PERSON**
Apple Inc.	**ORG**ANISATION
June 8th, 2023	**DATE**

Table 1.2 TODO: Textual data with according labels.

Unsupervised learning The unsupervised learning method discovers patterns of entity occurrence in raw and unlabeled data. Consequently, the individual entities are split into groups based on their characteristic properties. The absence of pre-labeled data by human intervention in the training phase causes no supervisor to guide the model with information about the labels in training data, hence unsupervised learning. A conventional method employed in this approach is clustering, such as K-means clustering (Sinaga et al., 2020), which divides data[3] into groups based on similarity or dissimilarity.

[3] In our case, the data corresponds to tokens symbolising words.

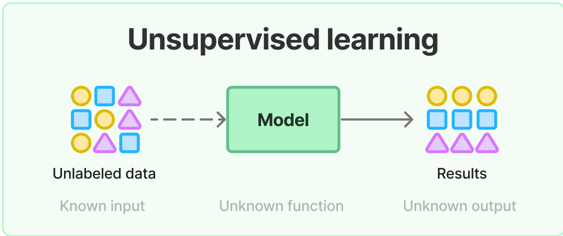

Figure 1.6 TODO: Unsupervised learning.

Semi-supervised learning The semi-supervised learning method combines labeled and unlabeled data, with the former comprising a slight portion of the dataset. As part of the classification training process, the unlabeled data learn the model's ability to generalise and represent the data in space using a statistical feature that better separates the classes. The algorithm aims to create the best decision boundary between classes based on a large amount of unlabeled data. Besides, the labeled data allows the model to determine the classification correctness for improvement, hence semi-supervised learning. Therefore, the model is partially supervised, hence semi-supervised learning. Semi-supervised learning is a preferred approach for model development because the labeled data is mainly expensive and time-consuming to acquire by requiring human intervention. In contrast, unlabeled data is more easily collectable.

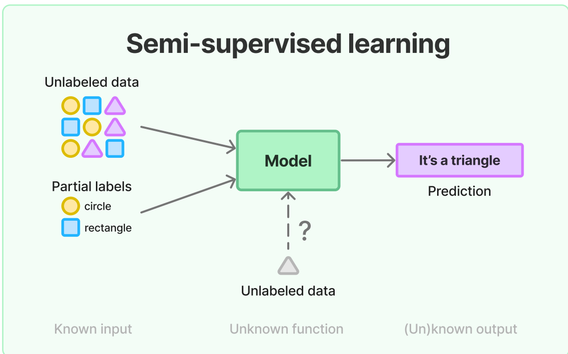

Figure 1.7 Semi-supervised learning.

Self-training, Co-training, and Multi-view learning are frequently employed subcategories of this learning method. Self-training involves using a small amount of labeled data and a more significant amount of unlabeled data, repeatedly training the model on labeled data and using it to predict labels for unlabeled data. Co-training uses multiple independent models that communicate and share information to improve performance. Multi-view learning combines information from different sources or representations of data, such as combining textual data with metadata.

Supervised learning Unlike the semi-supervised learning method, the supervised learning approach exclusively utilises labeled data. Therefore, the training process is performed just on labeled data. We could perceive this approach as a function $f : X \mapsto Y$, denoted as f, mapping input X to output Y, where X and Y represent the input and output, respectively, known from the labeled data. Thus, as a supervisor guides the learning process, the model learns the most suitable way to map the input X to the input Y. The principal contrast between fully and semi-supervised learning is that in the former case, learning is done only over labeled data and in the latter case on a combination with unlabeled data. The most common algorithms utilised in supervised learning for named entity recognition include statistical models like Support Vector Machine (SVM) (Wang, 2005), Conditional Random Field (CRF) (Sutton et al., 2012), Maximum Entropy (ME) (Berger et al., 1996), and Hidden Markov Model (HMM) (Eddy, 1996).

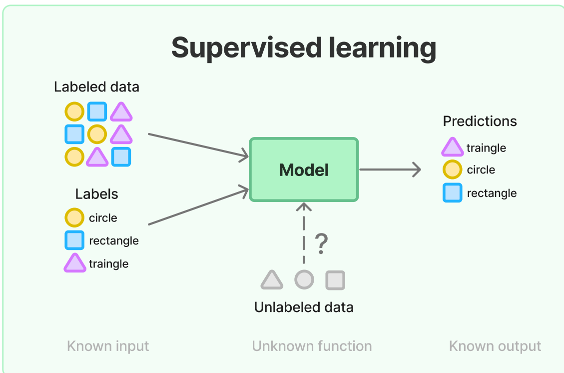

Figure 1.8 TODO: Supervised learning.

Deep learning

A given text could be extensive and contain many sentences or paragraphs mentioning entities in various forms. Therefore, deep learning comes into play due to its prowess in achieving better comprehension and learning contextual semantics. Its approaches overcome the primarily classification-focused methods introduced thus far, leading to state-of-the-art results in entity recognition. To enhance elucidation of the semantic context, methods' consideration of words occurring before and after words possess a role. The overarching structure of this approach's most commonly employed methods, such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Transformer, is usually outlined into three fundamental stages: Data representation, Context encoding and Entity decoding.

1. **Data representation** is the first stage, where the input data are transformed into a format suitable for the model. We convert words into vector space to operate effectively in the following stages and maximise the potential of computational power. Transformations, collectively referred to as word embeddings, encompass techniques like One-Hot Encoding, Word2Vec (Mikolov et al., 2013), and Term Frequency-Inverse Document Frequency (TF-IDF) (Aizawa, 2003).

    (a) **One-Hot Encoding** consists of a vector representation of a word with a 1 occurring at its positional index in the dictionary and a 0 otherwise. The dictionary can be thought of as a sentence that is a sequence of words in a row.

    (b) **Word2Vec** projects words in a vector space that captures the meaning and relationships between other words. The projection into vector space allows the creation of a unique identification for each word in the corpus, with an emphasis on preserving syntactic and primarily semantic properties, making it possible to find synonyms after training on a large corpus.

    TODO?: Přidat zmínku o CBOW a Skip-gram?

    (c) **TF-IDF** as the name suggests, takes into account both the frequency of a word in a document and its occurrence across all documents in the corpus. Hence, a word's importance could be captured and then deduced whether it is an entity.

    Additional representations worth mentioning are FastText (Joulin et al., 2016), Global Vectors for Word Representation (GloVe) (Pennington et al., 2014), or Embeddings from Language Model (ELMo) (Peters et al., 2018). While these data representations may also find application in feature engineering-based methods, they must be met in the context of deep learning for effective textual data utilisation in subsequent processes.

2. **Context encoder** conveys the most crucial stage. Once we have the data in a suitable representation capable of capturing the context in which the word is found, the type of neural network selection to learn the model

contextual semantics of the input data comes next. As the deep learning methods approach is introduced, it is good to mention that "deep" refers to the depth of the layers in the neural network. Thus, the neural networks mentioned afterwards contain more than one hidden layer of neutrons, a layer that is neither an input nor an output layer. The articles (IBM, 2024b; IBM, 2024a) describe what neurons are and how deep neural networks differ from regular neural networks.

TODO!!! Následující část s
mi nelíbí a pravděpodobně
udělám ještě krátkou a
ukázkovou sekci o Neu-
ral networks. Protože jsou
využívány v Sentiment Ana
ysis, Named Entity Recogn
tion, Time Series Forecastin
Text Similarity

3. **Entity decoder** is the final stage, where the model identifies and classifies entities in the text based on word representation and extracted context information. Entity decoding has two flavours based on the way the output of the previous stage is processed.

   (a) **Classification models** such as softmax regression or multilayer perceptrons (MLP) can be used to assign labels to individual words in the text based on their representation and context.

   (b) **Sequential models** such as recurrent neural networks (RNNs) or transformers are able to sequentially predict entity labels for individual words in a text given their context and relationships with other words.

RNN

CNN

Transformer 1. **BERT**

Large language models

1.3 Workflow of Entity-level Sentiment Analysis

The sentiment analysis process can be divided into three main steps: data retrieval, preprocessing, and analysis. The following sections will discuss these steps in more detail.

1.4 Time Series Forecasting Integration

We will focus on integrating time series forecasting in the context of news.

1.5 Text similarity

XXXXX
(Q-13) Zeptat se - Bude třeba úvod do stock marketu? – odpověď: ANO, bude třeba krátké intro.

2. Related work

While examining online applications with similar characteristics, a common challenge arises from the need for more transparency regarding the techniques applied in sentiment analysis. Specifically, employed methods and the details of the models remain undisclosed. We acknowledge this limitation due to the proprietary nature of the software, mainly as it is an essential part of the business model. Hence, this chapter is organized into two primary sections.

The initial part provides an overview of existing applications. At the same time, the second section will concern the most recent and relevant research on predicting stock market behaviour using data mining techniques and news sentiment analysis. Additionally, it explores research studies on entity-level sentiment analysis over news articles and its application in the stock market.

2.1 Existing Application Overview

Sentiment analysis applications accessible to the public are typically based on investigating social network posts, with StockTwist[1] as a notable illustration. According to Reuters (2022), StockTwits in 2022 boasts more than six million registered users and one million monthly active users, underscoring its prominent user base. With the growing volume of social media contributions, it is difficult to determine which post will prompt action. In our case, these posts do not constitute highly relevant data since our interest is in news articles containing a more significant amount of information. Even if the selection of accounts is limited to informative sources, News organizations' social media accounts only link to their articles, usually accompanied by a headline or lead paragraph. These posts are not enough for our purposes, as we need the entire article to perform better context for entity-level sentiment analysis (discussed further in Chapter 3).

Q-2.1 Lze takto citovat "pouze" webový článek deníku? Url je moc velký a chtěl bych se vyhnout tomu, abych ho tlačil do footnote

2.1.1 Bloomberg Terminal

The only software similar to the one created is a module in the so-called Bloomberg Terminal[2] from Bloomberg L.P.. This software system provides investors with analytical tools over financial data, including sentiment analysis of news articles and posts on social network X, formerly known as Twitter. The cost of

[1] *https://www.stocktwits.com*
[2] *https://www.bloomberg.com/professional/solution/bloomberg-terminal/*

TO[illegible]O: a[illegible]ebo[illegible]e ?

a Bloomberg Terminal depends on the required specific features and services. A standard subscription typically amounts to approximately $24,000 per year. It is a very complex and powerful software, but it is not accessible to the general public.

Bloomberg, in its work (Cui et al., 2024), describes two types of sentiment analysis: story-level and company-level sentiment, utilizing a suite of Supervised Machine Learning (SML) techniques. Classification engines are trained on labeled datasets containing news articles and social media posts. Reportedly, the labeling process is based on the question:[3]

> *"If an investor having a long position in the security mentioned were to read this news or tweet, is he/she bullish, bearish or neutral on his/her holdings?"*

Once model training is completed, the models are employed to analyze recently published posts and articles associated with organizations, seeking distinctive sentiment signals related to the business and finance domain. As mentioned above, sentiment is divided into two levels:

Story-level Sentiment score value of articles and posts is calculated after arrival in real time. The calculation includes score and confidence, where the score has one of three options: positive, negative, or neutral, each described by a numerical value from the set $\{-1, 0, 1\}$. The confidence is defined by a value ranging from 0 to 1, demonstrating the intensity[4] of the sentiment. Hence, the story-level sentiment ranges from -1 to 1. For both, we get the following equation:

$$\textit{Story-level}_c^{\textit{Articles}} = S_c^a C_c^a, \quad a \in P(c) \tag{2.1}$$

$$\textit{Story-level}_c^{\textit{Posts}} = S_c^p C_c^p, \quad p \in P(c) \tag{2.2}$$

where a represents an article and p represents a post from $P(c)$, the set of published articles and posts referencing company c. S_c^a and S_c^p are the sentiment polarity scores of article a and post p that reference company c. C_c^a and C_c^p are the confidences of article a and post p that reference company c.

Company-level Sentiment score value is then calculated as the confidence-weighted average of the story-level sentiment scores, incorporating all relevant news articles and social media posts mentioning the company.

$$\textit{Company-level}_{c,t}^{\textit{Articles}} = \frac{\sum_{a \in P(c,T)} S_c^a C_c^a}{N_{c,T}^a}, \quad T \in [t_b, t] \tag{2.3}$$

[3]This information is difficult to verify due to the unavailability of the dataset.
[4]Probability of being positive, negative, or neutral

$$\textit{Company-level}^{Posts}_{c,t} = \frac{\sum_{p \in P(c,T)} S^p_c C^p_c}{N^p_{c,T}}, \quad T \in [t_b, t] \tag{2.4}$$

where a represents an article and p represents a post from $P(c, T)$, the set of published articles and posts referencing company c during period T. Period T is a time interval of length t_b to t, where t_b is the time constant of the beginning. $N^a_{c,T}$ and $N^p_{c,T}$ are the number of articles and posts referencing company c during period T. In this way, the company-level sentiment is calculated as follows:

$$\textit{Company-level}_{c,t} = \textit{Company-level}^{Articles}_{c,t} + \textit{Company-level}^{Posts}_{c,t} \tag{2.5}$$

Intraday Company-level sentiment score for news articles is recalculated every two minutes, utilizing an eight-hour rolling window. The sentiment score for social network posts is recalculated every minute, employing a 30-minute rolling window. Due to the previous definitions, we can express these by a simple substitution of t_b depending on the rolling window as follows:

$$\textit{Intraday Company-level}^{Articles}_{c,t} = \frac{\sum_{a \in P(c,T)} S^a_c C^a_c}{N^a_{c,T}}, \quad T \in [t-8, t] \tag{2.6}$$

$$\textit{Intraday Company-level}^{Posts}_{c,t} = \frac{\sum_{p \in P(c,T)} S^p_c C^p_c}{N^p_{c,T}}, \quad T \in [t-0.5, t] \tag{2.7}$$

$$\textit{Intraday Company-level}_{c,t} = \textit{Company-level}^{Articles}_{c,t} + \textit{Company-level}^{Posts}_{c,t} \tag{2.8}$$

Daily company-level sentiment scores are published every morning about 10 minutes before the market opens. The calculation is determined as a confidence-weighted average of sentiment scores derived from the story-level sentiments of news and social media posts over the past 24 hours as follows:

$$\textit{Daily Company-level}_{c,t} = \frac{\sum_{d \in P(c,T)} S^d_c C^d_c}{N_{c,T}}, \quad T \in [t-24, t] \tag{2.9}$$

where document d represents a news article or social media post, the sentiment polarity score of document d referencing company c is denoted as S^d_c, and the confidence associated with this reference is represented by C^d_c. The set $P(c, T)$ encompasses non-neutral documents referencing company c published within the last 24 hours. $N_{c,T}$ expresses the count of non-neutral documents referencing company c during period T. This approach is further explored in terms of the informational role of social media by Gu et al. (2020).

2.2 Predicting Stock Market Behaviour

Several approaches for predicting stock market behaviour and price trends have been proposed, utilizing sentiment analysis of financial news and historical stock prices. Several studies prove a strong correlation between financial news sentiment and stock prices (Li et al., 2014) (Wan et al., 2021). Due to the nature of unstructured textual data, predicting stock market behaviour is a challenging task.

Khedr et al. (2017) focuses on creating an effective model for forecasting future trends in the stock market, using sentiment analysis of financial news and historical stock prices. The model achieves more accurate results than previous works by considering different market and company news types combined with historical stock prices. The experiments utilize datasets from three companies: Yahoo Inc., Facebook Inc.[5], and Microsoft Corporation. The authors use well-known and informative news sources such as Reuters, The Wall Street Journal, and Nasdaq. The first step of sentiment analysis to get the text polarity using a Naive Bayes classifier was shown to achieve accuracy from 72.73% to 86.21%, while the second step, which combines news sentiment with historical prices, improved prediction accuracy up to 89.80%. Moreover, that is why we find this study motivating and inspiring.

2.3 Entity-level Sentiment Analysis

This section will discuss the most recent and relevant research conducted in entity-level sentiment analysis and its application in news articles, including named entity recognition over news articles.

Zhao et al., 2021 employed RoBERTa, a Robustly Optimized Bidirectional Encoder Representations from Transformers (BERT) Pretraining Approach (Liu et al., 2019b), to propose a sentiment analysis and entity detection strategy in financial text mining and public opinion analysis in social media. In the first step, sentiment analysis, mainly focusing on negative polarity, is performed. Then, entity detection is considered in different granularities with MRC, Machine Reading Comprehension (Liu et al., 2019a), or sentence-matching tasks. As a result, this study serves entity detection differently than traditional Named Entity Recognition. The authors claim that the proposed method outperforms traditional sentiment analysis and entity detection methods. The authors also emphasize the importance of the financial domain, where the sentiment of a single entity can significantly affect the stock price.

[5]Known as Meta since 2021

Named entity recognition plays a significant role in entity-level sentiment analysis, as discussed in Section 2.3. Vybírám články, které stojí za zmínku.

2.4 Mining dynamic Social Networks

Ještě si nejsem jistý, zda-li je tato sekce potřebná. XXXXX

This section will discuss interesting research Jin et al., 2012 dealing with mining dynamic social networks and its application in the stock market.

Bibliography

Piryani, R. et al. (2017). *Analytical mapping of opinion mining and sentiment analysis research during 2000–2015*. In: *Information Processing & Management* 53.1, pp. 122–150. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2016.07.001. URL: https://www.sciencedirect.com/science/article/pii/S030645731630245X.

Saunders, Danielle (2020). *Domain adaptation for neural machine translation*. PhD thesis. Apollo - University of Cambridge Repository. DOI: 10.17863/CAM.66458. URL: https://www.repository.cam.ac.uk/handle/1810/319335.

Liu, Bing (2022). *Sentiment analysis and opinion mining*. In: Springer Nature. Chap. 3, pp. 31–36.

Wankhade, Mayur et al. (2022). *A survey on sentiment analysis methods, applications, and challenges*. In: *Artificial Intelligence Review* 55.7, pp. 5731–5780. ISSN: 1573-7462. DOI: 10.1007/s10462-022-10144-1. URL: https://doi.org/10.1007/s10462-022-10144-1.

Mary, A. Jenifer Jothi et al. (2017). *Jen-Ton: A framework to enhance the accuracy of aspect level sentiment analysis in big data*. In: *2017 International Conference on Inventive Computing and Informatics (ICICI)*, pp. 452–457. URL: https://api.semanticscholar.org/CorpusID:44112128.

Wang, Yequan et al. (2019). *Aspect-level Sentiment Analysis using AS-Capsules*. In: *The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: Association for Computing Machinery, 2033–2044. ISBN: 9781450366748. DOI: 10.1145/3308558.3313750. URL: https://doi.org/10.1145/3308558.3313750.

Rønningstad, Egil et al. (Oct. 2022). *Entity-Level Sentiment Analysis (ELSA): An Exploratory Task Survey*. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 6773–6783. URL: https://aclanthology.org/2022.coling-1.589.

Liu, Bing (2015). *Aspect and Entity Extraction*. In: *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 137–188.

Zhang, Lei et al. (2014). *Aspect and Entity Extraction for Opinion Mining*. In: *Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities*. Ed. by Wesley W. Chu. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–40. ISBN: 978-3-642-40837-3. DOI: 10.1007/978-3-642-40837-3_1. URL: https://doi.org/10.1007/978-3-642-40837-3_1.

Ramshaw, Lance et al. (1995). *Text Chunking using Transformation-Based Learning.* In: *Third Workshop on Very Large Corpora.* URL: https://aclanthology.org/W95-0107.

Keraghel, Imed et al. (2024). *A survey on recent advances in named entity recognition.* arXiv: 2401.10825 [cs.CL].

Sinaga, Kristina P. et al. (2020). *Unsupervised K-Means Clustering Algorithm.* In: *IEEE Access* 8, pp. 80716–80727. DOI: 10.1109/ACCESS.2020.2988796.

Wang, Lipo (2005). *Support vector machines: theory and applications.* Vol. 177. Springer Science & Business Media.

Sutton, Charles et al. (2012). *An introduction to conditional random fields.* In: *Foundations and Trends® in Machine Learning* 4.4, pp. 267–373.

Berger, Adam et al. (1996). *A maximum entropy approach to natural language processing.* In: *Computational linguistics* 22.1, pp. 39–71.

Eddy, Sean R (1996). *Hidden Markov models.* In: *Current Opinion in Structural Biology* 6.3, pp. 361–365. ISSN: 0959-440X. DOI: https://doi.org/10.1016/S0959-440X(96)80056-X. URL: https://www.sciencedirect.com/science/article/pii/S0959440X9680056X.

Mikolov, Tomas et al. (2013). *Efficient estimation of word representations in vector space.* In: *arXiv preprint arXiv:1301.3781.*

Aizawa, Akiko (2003). *An information-theoretic perspective of tf–idf measures.* In: *Information Processing & Management* 39.1, pp. 45–65. ISSN: 0306-4573. DOI: https://doi.org/10.1016/S0306-4573(02)00021-3. URL: https://www.sciencedirect.com/science/article/pii/S0306457302000213.

Joulin, Armand et al. (2016). *Bag of Tricks for Efficient Text Classification.* arXiv: 1607.01759 [cs.CL].

Pennington, Jeffrey et al. (Oct. 2014). *GloVe: Global Vectors for Word Representation.* In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Ed. by Alessandro Moschitti et al. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

Peters, Matthew E. et al. (2018). *Deep contextualized word representations.* arXiv: 1802.05365 [cs.CL].

IBM (2024b). *Neural Networks.* URL: https://www.ibm.com/topics/neural-networks (visited on 03/08/2024).

— (2024a). *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks.* URL: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/ (visited on 03/08/2024).

Reuters (2022). *After crypto, Stocktwits takes aim at stocks trading.* URL: https://www.reuters.com/technology/after-crypto-stocktwits-takes-aim-stocks-trading-2022-07-19/.

Cui, Xin et al. (2024). *Embedded Value in Bloomberg News & Social Sentiment Data.* Received via email from Bloomberg L.P. on 2.2.2024. URL: https://data.bloomberglp.com/promo/sites/12/725454457_EDFSentimentWP.pdf.

Gu, Chen et al. (2020). *Informational role of social media: Evidence from Twitter sentiment.* In: *Journal of Banking & Finance* 121, p. 105969. ISSN: 0378-4266. DOI: https://doi.org/10.1016/j.jbankfin.2020.105969. URL: https://www.sciencedirect.com/science/article/pii/S0378426620302314.

Li, Xiaodong et al. (2014). *News impact on stock price return via sentiment analysis.* In: *Knowledge-Based Systems* 69, pp. 14–23. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2014.04.022. URL: https://www.sciencedirect.com/science/article/pii/S0950705114001440.

Wan, Xingchen et al. (2021). *Sentiment correlation in financial news networks and associated market movements.* In: *Scientific Reports* 11.1, p. 3062. ISSN: 2045-2322. DOI: 10.1038/s41598-021-82338-6. URL: https://doi.org/10.1038/s41598-021-82338-6.

Khedr, Ayman E et al. (2017). *Predicting stock market behavior using data mining technique and news sentiment analysis.* In: *International Journal of Intelligent Systems and Applications* 9.7, p. 22.

Zhao, Lingyun et al. (2021). *A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts.* In: *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 1233–1238. DOI: 10.1109/CSCWD49262.2021.9437616.

Liu, Yinhan et al. (2019b). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* arXiv: 1907.11692 [cs.CL].

Liu, Shanshan et al. (2019a). *Neural Machine Reading Comprehension: Methods and Trends.* In: *Applied Sciences* 9.18. ISSN: 2076-3417. DOI: 10.3390/app9183698. URL: https://www.mdpi.com/2076-3417/9/18/3698.

Jin, Yingzi et al. (2012). *Mining dynamic social networks from public news articles for company value prediction.* In: *Social Network Analysis and Mining* 2.3, pp. 217–228. ISSN: 1869-5469. DOI: 10.1007/s13278-011-0045-5. URL: https://doi.org/10.1007/s13278-011-0045-5.

Acronyms

BERT Bidirectional Encoder Representations from Transformers. 16

BIO Beginning-Inside-Outside. 8

IO Inside-Outside. 8

IOB Inside-Outside-Beginning. 8

NLP Natural Language Processing. 4, 7

SML Supervised Machine Learning. 14