SEMINAR PAPER

Sentiment Correlation in Financial News Networks

written under the guidance of

Univ.Prof. Dipl.-Ing. Dr.techn. Stefan Gerhold

by

Amina Askarbek

Matriculation Number: 11732289

Vienna, on February 28, 2022

Contents

1 Introduction **1**

2 Sentiment Evolution **2**
2.1 Natural Language Processing . . . 3
2.1.1 Named Entity Recognition . . . 3
2.1.2 Convolutional Neural Networks . . . 4
2.1.3 Long Short-Term Memory . . . 5
2.1.4 Conditional Random Fields . . . 6
2.2 Entity Detection and Classification . . . 7
2.3 Sentiment Value and Evolution . . . 8

3 News Co-occurrence Network **10**
3.1 The Network . . . 10
3.1.1 Community Detection . . . 10
3.1.2 Visual Representation . . . 12
3.1.3 Outliers . . . 13
3.2 Centrality Measures . . . 14
3.2.1 Eigenvector Centrality . . . 14
3.2.2 Betweenness Centrality . . . 16
3.2.3 Comparison . . . 16
3.3 Network Movement . . . 17
3.3.1 Normalized Mutual Information . . . 17
3.3.2 F_1 Measure . . . 18
3.3.3 News Co-occurrence Network over Time . . . 18

Bibliography **20**

1 Introduction

This seminar paper is based on the article "Sentiment Correlation in Financial News Networks and Associated Market Movements" by X.Wan et al. (1).

Modern financial markets have a highly inter-dependent nature. The relations among participating entities on a financial market exhibit an interesting research ground, as the cohesiveness and mutual dependency of market participants may be studied to reveal underlying reciprocal influences of unexpected and expected nature. Whenever a news articles touches upon one company, some surprising inter-connections might surface. Additionally, the information about multiple companies being mentioned in news articles collectively can be crucial in the understanding of the underlying relationships between said companies.

In order to study possible relations among companies on a financial market, data-driven approaches are applied. Natural language processing techniques assist in the understanding of the connections in a financial market, and a construction of the news co-occurrence network allows us to examine the dynamics of sentiments and the market as a whole.

In this paper, methods of linguistic analysis are discussed, followed by the overview of an application of such methods, and the derived results. Moreover, the techniques used for the construction of the news co-occurrence network are reviewed, and the following analysis of the network is presented.

2 Sentiment Evolution

The development in the fields of data analysis and computational linguistics allow us to take a profound look on the possibility of discovery of relations that may be present on the financial market among various entities, which may not be recognized upon a primarily superficial look.

Although the news articles and other various texts which use human-spoken languages are recognized by us, it nevertheless is tricky to convert the information efficiently to be recognized by technological processes. While human languages can follow a predefined set of grammatical rules, the aspect of emotions and linguistic techniques that enrich the spoken and written language can be difficult to capture and precisely interpret by a computational algorithm.

One of the ways an article, or any other text source, can be analyzed is by the implementation of natural language processing (further abbreviated as NLP) techniques. The natural language processing techniques allow us to closely examine human languages computationally almost without losing the crucial aspect of emotion and subjectivity of a language. The numerical processing using NLP techniques is therefore an outstanding advancement in the collection and processing of linguistic data.

Once a language is processed efficiently enough to be understood by an algorithm, the question of further usage of the numerical data arises. As it is not enough to simply collect the data, the latter stage of manipulation of said data is critical to derive an appropriate analytical result.

As an example, the analysis of the linguistic data can be processed to demonstrate sentiments towards certain entities on the financial market. The underlying relationships of companies can be studied to group the entities in clusters and research the effect of a certain sentiment on a particular company, or on a cluster of related entities as a whole. Therefore, the understanding and numerical analysis of inter-dependencies on a financial market and of corresponding sentiments can be of key importance for further financial market research.

In order to analyse financial news texts derived from Reuters[1], a named entity recognition system, a character-level convolutional neural network with a word-level long short-term memory network, and a conditional random field were used. (1) An introduction to the aforementioned techniques is provided in the corresponding subsections.

[1] https://www.reuters.com

2.1 Natural Language Processing

The terminology "natural language" is used to describe the human-language. The defining feature of a natural language is that it hasn't been artificially manufactured, but rather developed on its own throughout the generations of human populations.(2) German, English, Spanish – would all fall under the definition of a natural language, whereas a programming language would be distinctly separated from the category of natural languages.

Since the field of linguistics, and more specifically computational linguistics, is an ever evolving part of contemporary research, there is a lack of a unifying definition for Natural Language Processing. The following definition shall provide sufficient detail to describe NLP in a mostly consistent and proper manner as presented by E.Liddy:

Definition 2.1.1. (3) Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

There are numerous NLP techniques that are being applied in linguistic analysis. Therefore, in the aforementioned definition there is no strict boundary set for a description of a unique and singular NLP technique. The definition also mentions *naturally occurring texts*, which are described to be of any nature and style, as long as the texts have not been adjusted to specifically accommodate the analysis, and can be found in a naturally occurring use of human language.(3)

2.1.1 Named Entity Recognition

One of the natural language processing techniques than has been used in this research is NCRF++, a named entity recognition (NER) system based on state-of-the-art deep learning.(4; 1)

Named Entity Recognition is a process that allows information from the text to be extracted and systematized. The *entities* are certain categories, for example *Person*, *Organization*, *Geopolitical Entity (GPE)*, and the unorganized text can be mined for words or phrases which fit the predefined entities. Such extraction and classification of information from a text has a wide range of applications in various scientific areas.(5)

The identification and classification of names and phrases in a text might not be a problem for humans, as we, due to our experience and awareness, are able to recognize whenever a certain name falls under a certain category. However, a machine can have a hard time. Natural language has its ambiguities and subjectivity, as for example *"Aurora"* can fall under the category *mythological name*, or under the category *natural event*, or it can also be under *person name*. The same way, without provided context, *Vienna University of Technology (TU)* can be categorized as both *place* or as *organization*. In the phrase *"How can I get to the Vienna University of Technology?"*, *Vienna University of Technology* falls under the

category *place*, but in the phrase *"Vienna University of Technology received an award"*, *TU* can be classified as *organization*.

NER operates in two phases, first being the recognition of certain phrases and words, and the second being the appropriate classification of the recognized elements. There are numerous methods of named entity recognition application, such as rule-based, machine learning-based, or hybrid-based NERs.(5)

Rule-based NER is a model that uses a set of predefined patterns that have to be inputted manually. The construction of a set of rules that complies with grammatical, lexical, orthographic, and syntactic features of a certain language is a tedious and sophisticated process that requires extensive expertise and accuracy.(5; 6) Moreover, the developed system, no matter how efficient and sturdy, cannot guarantee steady results in the long run.(6)

On the other hand, machine learning-based approach uses statistical models to accurately classify the identified elements. Machine learning used for named entity recognition can be split into two categories, the supervised and the unsupervised models. Supervised learning bases the mechanism of learning on a prepared training data. The program learns to identify examples with certain labels, and is therefore in need of a human-made supervision, hence the name. The downside of a supervised model is the vast amount of data it requires to achieve accurate results. The second method, unsupervised learning, has no need of human intervention and can learn with no supervision. However, unsupervised learning is not common in NER. (5)

The Hybrid NER model, as the name suggests, uses the advantageous features of both rule-based and machine learning-based systems. This model can deliver more favourable results, but the need for a human-crafted rule system remains its most prominent disadvantage.(5)

2.1.2 Convolutional Neural Networks

Convolutional Neural Network (CNN) is a deep learning algorithm which performs classifications in a text corpora. CNNs have shown to perform outstandingly in text systematization and are a key technique in natural language processing. For the purpose of text recognition, there are two types of convolutional neural networks: word-level CNN and character-level CNN.(7)

Word-level convolutional neural network is dependant on a trained word-model, which can cause issues if there is no such pre-trained model available at disposal. A word-model is moreover laborious and costly to configure and develop, as well as including a potential risk of misspellings and lack of certain lexicon in the predefined model. A character-level CNN, on the other hand, requires no trained model or any time-consuming pre-processing manipulation of the textual data. However, character-level CNNs are generally not as accurate as word-level CNNs.(7)

As the name suggests, convolutional neutral networks are based on the mathematical operation of convolution.(8)

Definition 2.1.2. Suppose f and g are two functions. The convolution $f * g$ is given by

$$(f * g)(t) := \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$$

In the CNN terminology, the function f would be called an *input*, and the function g would be a *kernel*.

We further define a discrete convolution, as occasionally the data might only have discrete values:

Definition 2.1.3. Discrete Convolution

$$(f * g)(t) = \sum_{\tau=-\infty}^{\infty} f(\tau)g(t-\tau)$$

Convolutional neural networks can process data that have a form of two dimensions (2D), such as images, or of one dimension (1D), such as text. Therefore, CCNs are useful for graphical image-processing and recognition, as well as textual analysis and characterization. The input and kernel data is usually represented by an array of multiple dimensions, called *tensors*, where in the point with no stored value, the array elements are zero.(8)

A simplified explanation of the architecture of a character-level CNN is as follows: first, the input text is fed into the algorithm. Then, a convolutional layer and a pooling layer are applied. After arbitrarily many repetitions of convolutional and pooling layers, an output is produced. The output is then a classification of the input text into various predefined categories.(9)

The pooling layers can use the module of max-pooling as defined by X.Zhang, J.Zhao, and Y.LeCun:

Definition 2.1.4. (9) Given a discrete input function $g(x) \in [1, l] \to \mathbb{R}$, the *max-pooling* function $h(y) \in [1, \lfloor (l-k+1)/d \rfloor] \to \mathbb{R}$ of $g(x)$ is defined as

$$h(y) = \max_{x=1}^{k} g(y \cdot d - x + c)$$

where $c = k - d + 1$ is an offset constant.

The max-pooling function is a preferred method in pooling layers as it enables training deeper than 6 levels.(9)

2.1.3 Long Short-Term Memory

Long Short-Term Memory is a recurrent neural network architecture, which solves the problem of back-propagated error signals that either vanish (tend to zero) or blow-up (tend to infinity) of conventional architectures such as *Back-Propagation Through Time* or *Real-Time Recurrent Learning*.(10)

2.1.4 Conditional Random Fields

The two common probabilistic models that are applied to extraction and labeling processes are Hidden Markov models (HMM) and stochastic grammars. Both models allocate a joined probability to paired observation and label sequencing, with the usual training of parameters in maximizing the joined likelihood of training samples. However, the models are not appropriate when used to represent several interacting features or observations with long-range dependencies. To avoid this issue faced with HMMs and stochastic grammars, we introduce Maximum entropy Markov models (MEMMs).(11)

MEMMs are conditional probabilistic models, which have certain advantages when compared to Hidden Markov models. A conditional model is given an observational sequence, which is then assigned a feasible label of a specified probability. It does not model the observations themselves, and the conditional probability of label assignment depends on arbitrary, dependent features of the observations. The model does not have to account for the dependencies' distributions. Moreover, in case past and future observations are available, the labeling transition probability can depend on those observations, as well as the current one. A generative model such as HMM, by contrast, is forced to take strict independence assumptions of the observational sequences.(11)

Maximum entropy Markov models have a weakness, the so called *label bias problem*: a mutual competition of transitions leaving a given state, rather than a competition against other transitions in the model. The *conditional random fields (CRFs)* have all the advantages of MEMMs, but lack the disadvantage of the label bias problem. Furthermore, while a MEMM's conditional probabilities (next stages given current stage) are based on per-state exponential models, a CRF's joint probability (all labels given all observations) is built on a sole exponential model, which is the main distinction between CRFs and MEMMs.(11)

A formal definition as given by J.Lafferty, A.McCallum, and F.Pereira:

Let $\mathbf{X}$ be a random variable over data sequences to be labeled, and let $\mathbf{Y}$ be a random variable over corresponding label sequences. The random variables $\mathbf{X}$ and $\mathbf{Y}$ are jointly distributed with a conditional model $p(\mathbf{Y} \mid \mathbf{X})$, but we do not explicitly model the marginal $p(\mathbf{X})$.

Definition 2.1.5. (11) Let $G = (V, E)$ be a graph such that $\mathbf{Y} = (\mathbf{Y}_v)_{v \in V}$, so that $\mathbf{Y}$ is indexed by the vertices of G. Then $(\mathbf{X}, \mathbf{Y})$ is a *conditional random field* in case, when conditioned on $\mathbf{X}$, the random variables $\mathbf{Y}_v$ obey the Markov property with respect to the graph:

$$p(\mathbf{Y}_v \mid \mathbf{X}, \mathbf{Y}_w, w \neq v) = p(\mathbf{Y}_v \mid \mathbf{X}, \mathbf{Y}_w, w \sim v)$$

where $w \sim v$ means that w and v are neighbors in G.

2.2 Entity Detection and Classification

Using the aforementioned natural language processing techniques, financial news articles on Reuters were analysed. The goal of such text analysis was to extract entities (i.e., companies) from unstructured news texts. The time period of the selected news articles ranges from 1st of January 2007 to 30th of September 2013, which makes up 27 full quarters.

Applying the NLP techniques, the entities classified from the financial texts are kept if they fall under the category *Organization*. Moreover, a company can be referred to using different name formats, as for example "IBM" can also be mentioned as "International Business Machines Corporation". Therefore, to sort out entities referring to the same organization, a couple of methods are applied:

1. hand-written rules to identify various name formats of the same organization;
2. automatic disambiguation of entities through brackets.

To elaborate on the second method, whenever a news article mentions a full name of an organization with an abbreviation in brackets, the information is used to imply that the abbreviation and the full name refer to the same company. For example, if the news article states: "American Express Co (AXP.N) said on Thursday...", *American Express Co* and *AXP.N* would be interconnected as both reference to the same organization.

The organizations that are mentioned consistently more than four times in each of the total 27 quarters are kept, with the number of frequent organizations totalling 145 as a result. However, the companies are processed further depending on the relevant price ticker in the Bloomberg terminal, resulting in 87 final organizations. The categorization of the 87 relevant companies results in the establishment of nine sectors, which are based on the Bloomberg sector list of 2018.

The sectors and the total number of companies included in each sectors, with a couple of examples of companies in the sector, are represented in the following table.

Sector	Number of companies	Examples
Financials	25	JP Morgan, Deutsche Bank
Technology	16	Apple, Google, Moody's inc.
Materials	4	Alcoa, Inspiration Mining
Energy	5	Chevron, Ecopetrol
Communications	2	Verizon Communications, Vodafone
Healthcare	2	Johnson & Johnson, Pfizer
Consumer Staples	6	Pepsi, Walmart, Target Corp.
Consumer Discretionary	20	Amazon, Delta Airlines, Starbucks
Industrials	7	Airbus SE, Boeing, FedEx

2.3 Sentiment Value and Evolution

For the 87 final companies the sentiment value was computed using a state-of-the-art sentiment prediction algorithm. The value of sentiment plays an import role in further analysis of the sentiment development and market dynamics. In order to quantify a sentiment of an organization, a scale from -1 to +1 is introduced, where -1 represents negative sentiment, 0 neutral, and +1 positive sentiment. The full financial news text is analysed to extract information which helps quantify the sentiment of the news for the targeted entity. Such extraction is done by a classical NLP method of *targeted sentiment analysis*.(12)

Figure 2.1 represents the average news sentiment score during the observed period of 27 quarters. The matrix is color-coded depending on the sentiment score, with red being a color of positive sentiment, white neutral, and blue coloring negative sentiments.

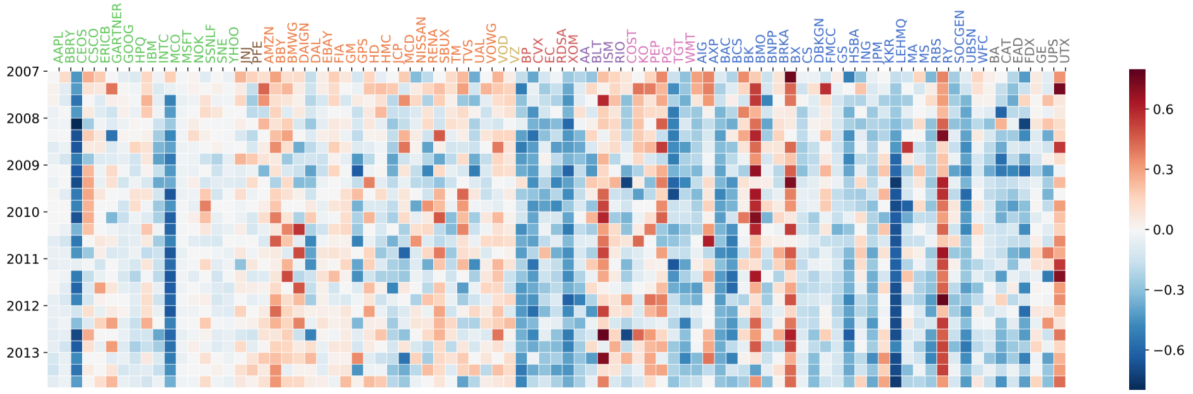


Figure 2.1: (1) Sentiment scores of 87 analyzed companies during a time period of 27 quarters

U.S. banks, e.g., Lehman Brothers (LEHMQ), JP Morgans (JPM), and Goldman Sachs (GS), have an observably strong negative sentiment during the financial crisis in September 2008, a clear indication of the impact of the crisis on those banks. On the other hand, the Canadian banks, such as Bank of Montreal (BMO) and Royal Bank of Canada (RY) do not seem to be impacted by the financial crisis, as their sentiment score stays positive despite the financial developments of that time. This, in fact, coincides with the general attitude towards Canadian banks during the period.

Moreover, there is a need to average out the sentiment values, since an entity might be mentioned in the news articles multiple times during an arbitrary period of analysis (such as, during a given day or during a span of a week). In case a sentiment is non-neutral, the news text carrying this sentiment is considered to be sentiment-bearing. For each of the companies, a sentiment score is therefore produced, during arbitrarily chosen time periods of evaluation. Furthermore, in order to compute a sentiment score across all sectors, the sentiment value of each company in a sector is aggregated and averaged.

In the following Figure 2.2, the average sentiments of each sector are presented.

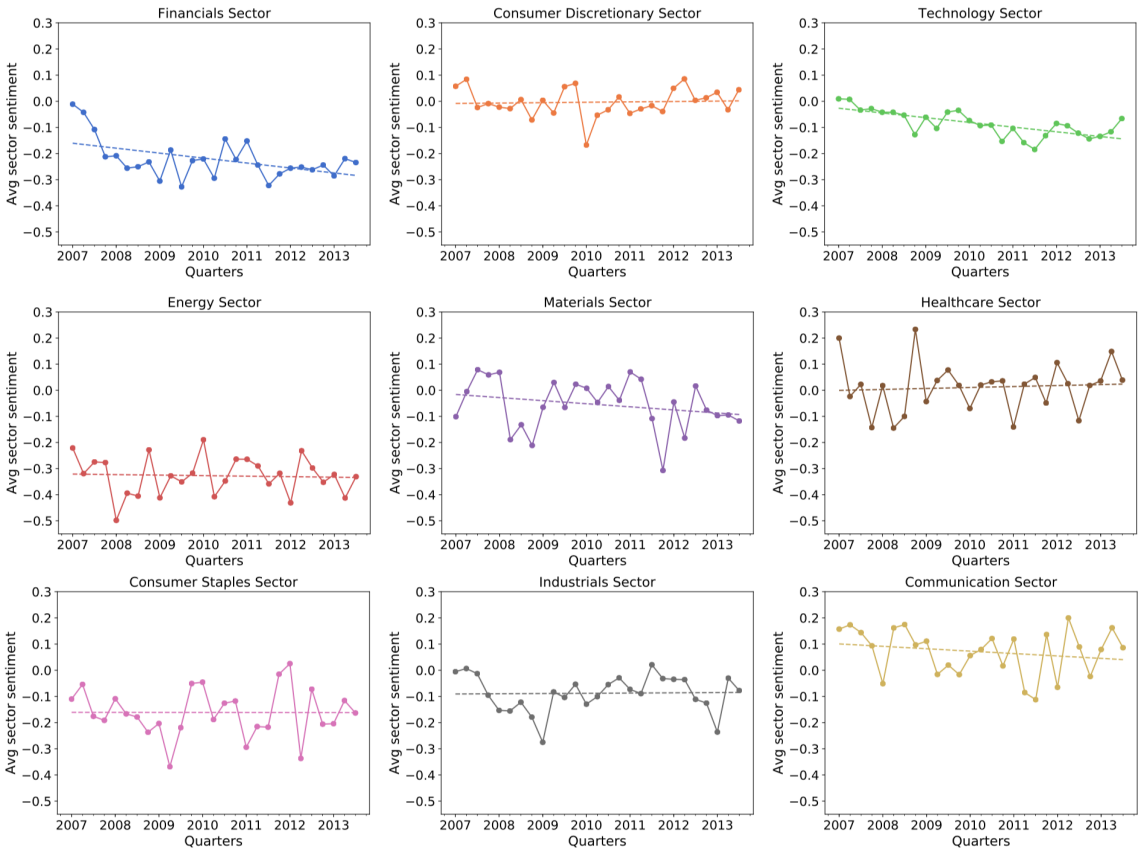


Figure 2.2: (1) Average sentiment of the nine sectors

It is noticeable that across the investigated quarters the entities in the sector *Technology* had a relatively steady sentiment development. However, the *Financial Services* sector tends to be turbulent. During the financial crisis quarters most sectors were rather volatile, which is a direct indication of the impact of the financial crisis. Additionally, the mostly negative sentiment score of the sectors *Financial Services* and *Energy* coincides with the individual sentiment values of the companies in the aforementioned sectors.

3 News Co-occurrence Network

3.1 The Network

The companies on the financial market are astonishingly inter-connected, the information about co-appearing companies in news articles can be used to investigate the underlying relationships between them. To research the co-dependency, a news coverage matrix is constructed.

In a news coverage matrix, the companies are allocated as rows, and the corresponding news are placed as columns. Let i be the rows, i.e., the companies, and let j be the index of the columns, i.e., the news. The ij-th element of the news coverage matrix represents the amount of news articles company i is mentioned in news j.

Using the first year of the collected data, year 2007, a weighted news co-occurrence network is designed. The nodes in the network embody the companies, which are connected in case of a relationship between them. Such connection is given by the edges of the network. Moreover, the edges are weighted with $e_{i,j}$, which is defined by the cosine distance of the corresponding row vectors in the matrix between the nodes i and j.

Definition 3.1.1 (Cosine Similarity)**.** Let $\mathbf{x}$ and $\mathbf{y}$ be two non-zero vectors. The *cosine similarity* is given by:

$$S_C(\mathbf{x}, \mathbf{y}) \coloneqq \cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

where x_i and y_i are components of the vectors $\mathbf{x}$ and $\mathbf{y}$ respectively, and θ the angle between the two vectors $\mathbf{x}$ and $\mathbf{y}$.

Definition 3.1.2 (Cosine Distance)**.** The *Cosine distance* is given by:

$$D_C(\mathbf{x}, \mathbf{y}) \coloneqq 1 - S_C(\mathbf{x}, \mathbf{y})$$

3.1.1 Community Detection

Informational networks can be complex in their structure, which complicates proper analysis of the connections of the elements in those networks. As a solution to ease the process a de-

composition of sophisticated networks into simpler communities was proposed. Such communities not only help in the reconstruction of the source network, but might also reveal hidden inter-connectivity within the network. This process is named *community detection*.(13)

The communities of a decomposed network consist of densely inter-connected nodes, with infrequent connections outside the detected communities. On the News Co-occurrence Network, community detection is used to obtain clusters of inter-connected companies. There exist multiple community detection algorithms (13), such as:

- divisive, which detects and removes inter-community links;
- agglomerative, which recursively merges similar nodes and communities;
- optimizing, which maximizes an objective function.

The latter type of community detection is applied in this research, in particular, the *Louvain modularity* method. The algorithm is performed to detect communities of related companies which we further name as *groups*.

Definition 3.1.3 (Louvain modularity method)**.** (1) The method uses *modularity*[1] as the objective function Q it aims to maximize:

$$Q = \frac{1}{2m} \sum_{i,j} \left(e_{ij} - \frac{k_i k_j}{2m} \delta(C_i, C_j) \right),$$

where the sum is over all edges in the network, the weight $e_{i,j}$ of the edge connecting two nodes i and j, the sum of all weights k_i and k_j of the edges i, j respectively, the communities C_i and C_j to which i and j belong to respectively, and δ the Kronecker delta function,

$$\delta(C_i, C_j) = \begin{cases} 1 & \text{if } C_i = C_j, \\ 0 & \text{otherwise.} \end{cases}$$

Initially, the Louvain method assigns a particular community to each node. The change in modularity ΔQ is computed by removing the node i from its assigned community and moving it consequently into every other community, to which the node i is connected to. When the process is completed for all communities, a community with the largest change in modularity ΔQ is established, and the node i is reallocated to this community. The latter step only occurs if a change in modularity is possible, otherwise the node i does not get reallocated. The whole process is then repeated successively until all nodes return no further change $\Delta Q > 0$.

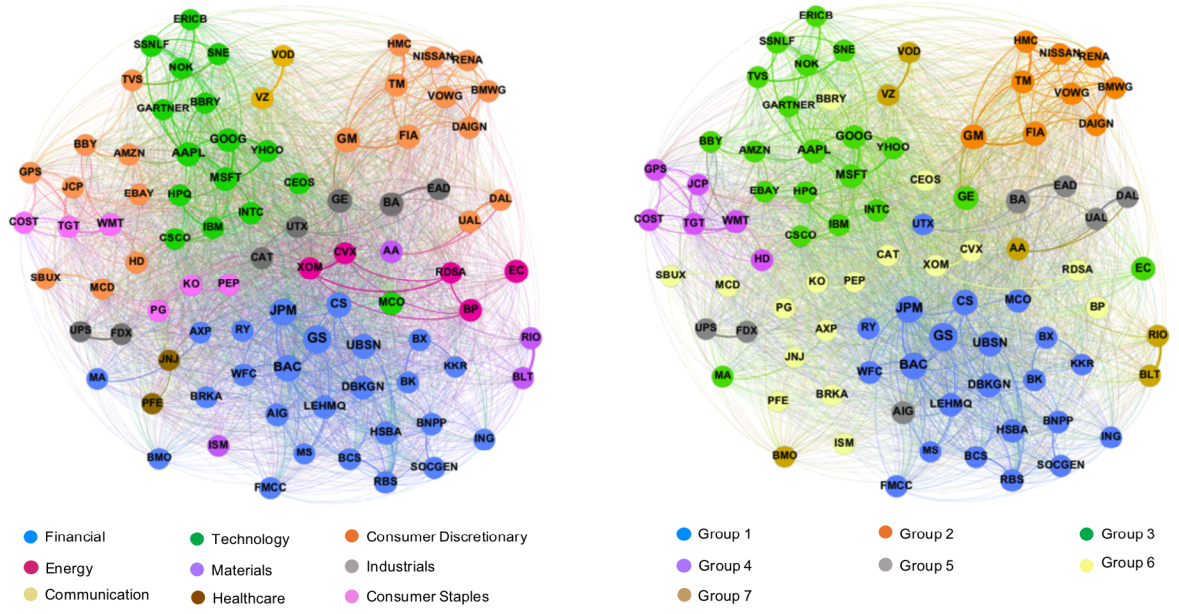
Upon conclusion of the aforesaid mechanism, the second phase initiates. During this step, the nodes that belong to the same community are considered as nodes of a new network, on which the first phase of the algorithm is reapplied once again. The second phase, and therefore the whole algorithm, terminates whenever no further change in communities is detected.

[1] A measure of the structure of networks

3.1.2 Visual Representation

As a result of the application of the community detection algorithm described in the previous section, a total number of seven communities (or, as previously defined, *groups*) was identified. The number is comparable to the nine Bloomberg-based sectors explained in the previous Chapter 2. Since some sectors, e.g., *Communications* and *Healthcare*, are rather small, we cannot expect exactly the same number of groups, as it would lead to sub-optimal segmentation.

The following Figure 3.1 presents the constructed networks, where Figure 3.1a is a visualization of the Bloomberg-based sectors, and Figure 3.1b is the representation of the groups yielded from community detection.



(a) Bloomberg defined sectors

(b) Detected groups

Figure 3.1: (1) News Co-occurrence Networks

If we compare the two networks, we can notice quite a bit of consistency and coherence between the sectors and the groups. In Figure 3.2 the detailed distribution of sectors into groups is given.

Additionally, although the median weight of an edge between two companies, which do not come from the same sector (otherwise called *out-sector*) is 0.00229, the median weight of an edge between two companies that belong to the same sector (known as *in-sector*) is 0.0157. This indicates that the companies which are often mentioned together in the news articles are more likely to belong to the same sector.

The Figure 3.3 visualizes the probability of weight distribution between *in-sector*, *out-sector* and *all* companies. The dotted vertical line represents the population median for each type

of companies. The median of *out-sector* and *all* overlay, since the majority of companies is allocated to be *out-sector*.

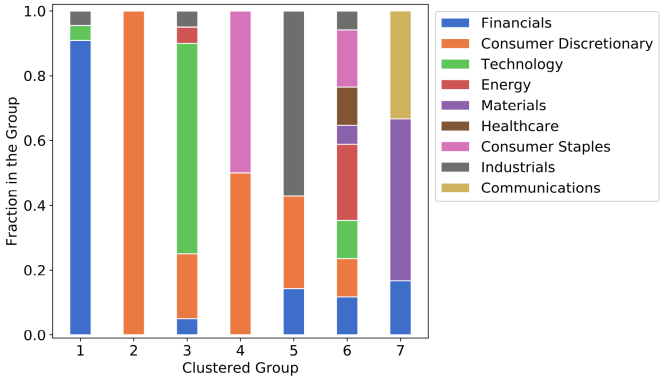


Figure 3.2: (1) Distribution of sectors in groups

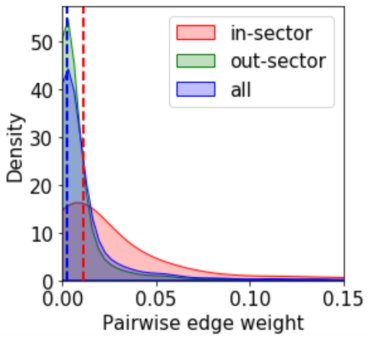


Figure 3.3: (1) Probability of weight distribution

Upon a closer examination of the detected groups and the Bloomberg sectors, the composition of the sectors is observed. As illustrated in Figure 3.2, Groups 1 and 2 almost entirely consist of sectors *Financials* and *Consumer Discretionary* respectively. In fact, an even more fine-grained information is revealed, as Group 2 mostly consists of *Automobile Manufacturers*, which is a sub-sector of the larger *Consumer Discretionary* sector. Contrariwise, Groups 3, 4, and 5, comprise of numerous sectors. This is justified by companies such as eBay or Amazon, which belong to the sector *Consumer Discretionary*, but nevertheless are combined with other companies that are classified as *Technology*. Group 4 is split between *Consumer Discretionary* and *Consumer Staples*, consisting mostly of conventional retail companies. Groups 7 and 8 consist of various smaller sectors that could not be assigned to a separate category.

3.1.3 Outliers

We might wonder, whether this analysis can reveal more fine-grained information about the relations of companies. Companies that belong to the same sector and exhibit particularly strong interconnections, or companies that have a significant relationship being parts of different sectors might be of interest.

To determine which company pairs can be considered outliers, a statistical approach is implemented. Namely, a pair of companies is considered to be an outlier, if the edge weight is above the 75th percentile + 1.5 Interquartile Range (IQR). This statistical approach is applicable to both groups, i.e., out-sector companies that exhibit interconnections, and in-sector company pairs with a considerably strong relation.

In the following tables, snippets of the results are presented with the most interesting company pairs from both categories. The *Rank* given in the tables indicates the positions of

largest edge weights between the paired companies in descending order, e.g., Rank 1 would be a position with the largest edge weight.

In-sector		Out-sector	
Rank	Company Pair	Rank	Company Pair
1	Vodafone – Verizon	2	Gap Inc. – Target Corp.
3	Microsoft – Yahoo	3	Alcoa – Delta Airlines
4	FedEx – UPS	4	Daimler AG – Airbus SE
5	Nissan – Renault	6	Amazon – Google
8	Coca-Cola – Pepsi	10	Best Buy – Walmart

Specifically, a couple of examples are worth noting in both segments. In the category of in-sector company pairs, we can observe especially strong links between companies that are competitors, or companies with an overlay in business areas. An example of such pairs would be Vodafone – Verizon or Goldman Sachs – Morgan Stanley (Position 6).

In the other category of out-sector companies we note Amazon, which is classified in *Consumer Discretionary*, having a strong relation to Google and Apple, both categorized in *Technology*, and also connecting to retail companies, such as Walmart. This is explainable by the market position and operating model of Amazon. Another example would be the relationship of Berkshire Hathaway with Proctor & Gamble, and with Coca-Cola, which is interpreted by the large share Berkshire Hathaway holds for both Proctor & Gamble and Coca-Cola.

3.2 Centrality Measures

From the observed results, we can draw a conclusion that the news co-occurrence network can provide a deeper insight into the relations between companies, which might not always be revealed upon a first look. The networks can identify company classes in a special manner that may not be available in traditional categorization. Moreover, the network can be constructed in a dynamic way, thus enabling information capture during an arbitrary period of time.

Inspired by the latter point, a news co-occurrence network for each quarter of the year is constructed for the purpose of analyzing the evolution of company position in the network. Such developments are investigated using the methods of eigenvector and betweenness centrality measures.

3.2.1 Eigenvector Centrality

Likewise to other centrality measures, eigenvector centrality assesses importance or rank of a node. Eigenvector centrality depends on the connectivity of a node (how many connections a node has to other nodes), and on the eigenvector centralities of other connected

nodes. Therefore, for the score of an arbitrary node, the contributions of nodes which have a higher eigenvector centrality will be more significant than the contributions of nodes with low eigenvector centrality.
The definition as given by M.Zaki et al.:

Definition 3.2.1. (14) Let $G = (V, E)$ be a direct graph, with $|V| = n$ vertices. The adjacency matrix of G is an $n \times n$ asymmetric matrix $\mathbf{A}$ given as

$$\mathbf{A}(u, v) = \begin{cases} 1 & \text{if } (u, v) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 3.2.2. (14) Let $p(u)$ be a positive real number, called the *eigenvector centrality* score of node u:

$$p(v) = \sum_u \mathbf{A}(u, v) \cdot p(u) = \sum_u \mathbf{A}^\intercal(v, u) \cdot p(u)$$

Across all nodes we can also express the eigenvector centrality scores as

$$\mathbf{p}' = \mathbf{A}^\intercal \mathbf{p}$$

where $\mathbf{p}$ is an n-dimensional column vector corresponding to the eigenvalue centrality scores for each vertex.

The centrality measure can be written in a vector notation as the eigenvector equation

$$\mathbf{A}\mathbf{p} = \lambda \mathbf{p}$$

where $\mathbf{p}$ is a vector of eigenvector centrality scores and λ an eigenvalue.

Eigenvector centrality can moreover be obtained recursively (14). If $\mathbf{p}_{k-1}$ is a vector of eigenvector centrality scores across all nodes at iteration $k-1$, then the vector of eigenvector centrality scores at iteration k is given by:

$$\begin{aligned} \mathbf{p}_k &= \mathbf{A}^\intercal \mathbf{p}_{k-1} \\ &= \mathbf{A}^\intercal (\mathbf{A}^\intercal \mathbf{p}_{k-2}) \\ &= \left(\mathbf{A}^\intercal\right)^2 (\mathbf{A}^\intercal \mathbf{p}_{k-3}) \\ &= \cdots \\ &= (\mathbf{A}^\intercal)^k \mathbf{p}_0 \end{aligned}$$

where $\mathbf{p}_0$ is the initial vector of the eigenvector centrality scores.

3.2.2 Betweenness Centrality

The betweenness centrality for a vertex v_i measures the number of shortest paths among all pairs of vertices which include v_i.

Definition 3.2.3. (14) Let η_{jk} be the number of shortest paths between vertices v_j and v_k, let $\eta_{jk}(v_i)$ be a number of paths that contain v_i. The fraction of paths through v_i is given by

$$\gamma_{jk}(v_i) = \frac{\eta_{jk}(v_i)}{\eta_{jk}}$$

If the vertices v_j and v_k are not connected, we assume $\gamma_{jk} = 0$.

Definition 3.2.4. (14) The betweenness centrality for a node v_i is defined as

$$c(v_i) = \sum_{j \neq i} \sum_{\substack{k \neq i \\ k > j}} \gamma_{jk} = \sum_{j \neq i} \sum_{\substack{k \neq i \\ k > j}} \frac{\eta_{jk}(v_i)}{\eta_{jk}}$$

3.2.3 Comparison

The two centralities were applied to the top six companies in the sectors *Financial* and *Technology*. The companies are Google (GOOG), Microsoft (MSFT), Apple (AAPL) from the sector *Technology*, and Goldman Sachs (GS), JP Morgan (JPM), Lehman Brothers (LEHMQ) from the sector *Financials*. The graphs of the centralities are given in Figure 3.4.

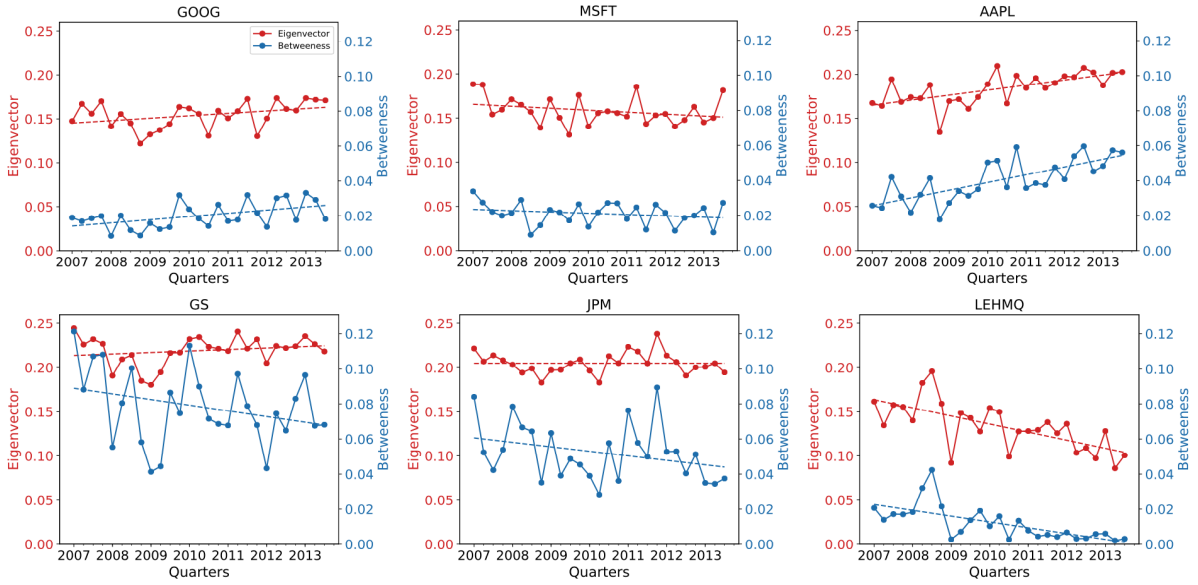


Figure 3.4: (1) Comparison of eigenvector and betweenness centralities

From the aforementioned graphs we can observe that both centralities of Google and Apple were increasing during the seven years, while Microsoft was decreasing. The centrality of all three financial companies was decreasing, and the centrality of Lehman Brothers faded after its bankruptcy during the financial crisis. Given this research, the dynamics of the company standings can therefore be looked at from an interesting perspective.

3.3 Network Movement

The evolution of the news co-occurrence network can be analyzed further by looking at the structure if the network over time. In order to do this, we have to take a look at two essential methods, the Normalized Mutual Information (NMI) and F1, as measures of group similarity.

3.3.1 Normalized Mutual Information

Mutual information (MI) is a measure of mutual dependence of two random variables. The observation of one random variable leads to information about another random variable, which is then quantified by the MI measure. There are multiple measures of mutual information, and hence, multiple normalized versions. We will take a look an one of the measures with its normalization.

Definition 3.3.1. (15) Let X and Y be two discrete random variables. The *mutual information* is given by

$$I(X;Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i, y_j) \ln\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right)$$

where $p(x_i, y_j)$, $p(x_i)$, and $p(y_i)$ are joint and marginal probabilities.

In the aforementioned definition, $I(X;Y)$ is a weighted mean of

$$I(x_i; y_j) = \ln\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right)$$

which is a measure of mutual information between two events $X = x_i$ and $Y = y_j$. This measure is not always non-negative, therefore another measure of the mutual information between two events is constructed

$$I(x_i; y_j) = \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \ln\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right) - \frac{p(x_i, y_j)}{p(x_i)p(y_j)} + 1 \geq 0$$

The normalization of this measure would then be defined in the following way.

Definition 3.3.2 (Normalized Mutual Information)**.** (15) Let X and Y be two random variables. Let the upper bounds of $I(X;Y)$ be $U_x = H(X)$ and $U_y = H(Y)$, where $H(X)$

and $H(Y)$ is the entropy of the random variables, defined for X, and in a similar manner for Y, as

$$H(X) = -\sum_{i=1}^{n} p(x_i) \ln p(x_i)$$

with $p(x_i)$ being the probability.
The *normalized mutual information* I^* is then given by

$$I^*(X;Y) = \frac{2I(X;Y)}{H(X) + H(Y)}$$

3.3.2 F_1 Measure

F-score measures are used to compute a test's accuracy. We define a couple of notions.

Definition 3.3.3. (16) Given the evaluation metrics that measure two different aspects of performance, recall and precision, we define

$$\text{recall} = \frac{\text{correct} + 0.5 \cdot \text{partial}}{\text{possible}}$$
$$\text{precision} = \frac{\text{correct} + 0.5 \cdot \text{partial}}{\text{actual}}$$

Recall is the percentage of possible answers that were correct. *Precision* is the percentage of actual answers given which were correct.

Definition 3.3.4. The F_1 *score* is the harmonic mean of precision and recall

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

3.3.3 News Co-occurrence Network over Time

The two aforementioned concepts were used to determine and render the group similarity.

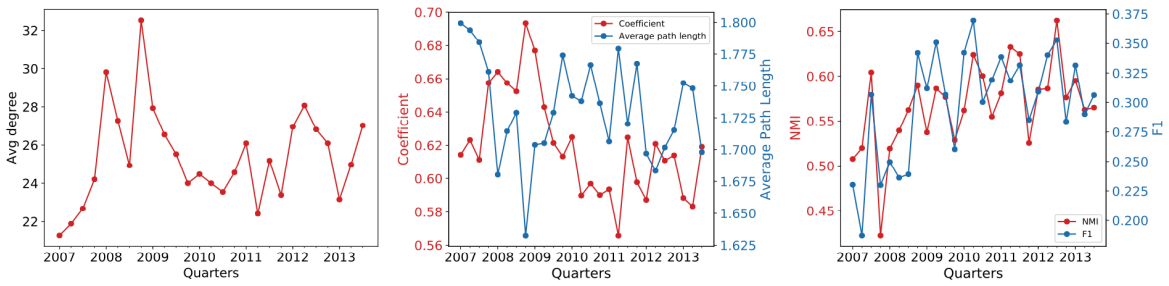


Figure 3.5: (1) Evolution of news co-occurrence network features and comparison with sectors

On the first graph in Figure 3.5 we can observe the development over quarters of the average degree of the news co-occurrence network. During the last quarter of 2008 the average degree of the network reached its peak, which might be explained by a rising number of companies being reported in the news together during the events of the escalation of the financial crisis and the failure of the market.

Shown on the second graph, the clustering coefficient and the average path length are a direct representation of the rising network interconnectedness.

And lastly, shown in the third graph are the NMI and F_1 measures of similarity between the derived groups and the Bloomberg-based sectors. It is noticeable that the similarity is mostly increasing, which might be explained by the increasing number of news articles over the observed period. The greater number of financial news allows us to derive the relations between companies more accurately.

Bibliography

[1] X. Wan, J. Yang, S. Marinov, J.-P. Calliess, S. Zohren, and X. Dong, "Sentiment correlation in financial news networks and associated market movements," *Scientific Reports*, vol. 11, no. 1, p. 3062, 2021.

[2] E. Kumar, *Natural language processing*. IK International Pvt Ltd, 2011.

[3] E. D. Liddy, *Natural Language Processing*. NY. Marcel Decker, Inc, 2 ed., 2001.

[4] J. Yang and Y. Zhang, *NCRF++: An Open-source Neural Sequence Labeling Toolkit*. 2018.

[5] A. Mansouri, L. S. Affendey, and A. Mamat, "Named entity recognition approaches," *International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 339–344, 2008.

[6] G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, "Using machine learning to maintain rule-based named-entity recognition and classification systems," in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 426–433, 2001.

[7] T. Londt, X. Gao, B. Xue, and P. Andreae, "Evolving character-level convolutional neural networks for text classification," 2020.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[9] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, p. 1735–1780, 1997.

[11] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[12] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification.," pp. 151–160, 01 2011.

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, Oct 2008.

[14] M. J. Zaki, W. Meira Jr, and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[15] T. O. Kvålseth, "On normalized mutual information: Measure derivations and properties," *Entropy*, vol. 19, no. 11, 2017.

[16] N. Chinchor, "Muc-4 evaluation metrics," in *Proceedings of the 4th Conference on Message Understanding*, MUC4 '92, p. 22–29, Association for Computational Linguistics, 1992.