The Austrian Start-Up Incubator Ecosystem

A Web Scraping, AWS ML & Text Analytics Competitor Analysis on Digital Content

By
Cosmin Catalin Ticu

Submitted to Central European University - Private University
Department of Economics and Business

In partial fulfilment of the requirements for the degree of
Master of Science in Business Analytics

Supervisor: Miklos Koren

Budapest, Hungary
June 8th, 2021

Executive Summary

This project is part of a larger effort by CEU's own iLab to run a full-scale competitor analysis on the Austrian start-up ecosystem as part of the iLab's expansion into the Vienna campus of the university. The goal of this capstone is to conduct an exploratory data analysis (EDA) on all of the proprietary visual and textual content curated by Austrian (with a keen focus on Vienna) start-up incubators on their own websites in order to identify formal patterns of association and common topics. The need for this project stems from the client's lack of explicit knowledge on the SEO and content creation efforts employed by its Austrian competitors and the client's need to diversify their content making efforts. The goal is to move from relying entirely on tacit knowledge, industry experience and networking acumen to employing a more data-driven approach for content creation. The steps taken to ensure a thorough EDA included: own definitions of methodology and strategy, self-gathering of data through web scraping, data munging and augmentation with AWS machine learning services and analysis of processed data with the R programming language to produce visualization artefacts.

The findings of this study echo the tacit content creation knowledge that "plagues" the entire start-up incubator market. By comparing all of the 819 articles available in the identified sample of 14 Austrian start-up incubators, this project found that the digital content produced is extremely similar throughout in terms of sentiment, key words, phrases, and image entities. There were rarely any elements found that set certain incubators apart from the group. The only distinctions that became apparent from the analysis were between specialized and non-specialized companies (here understood as having a narrow market focus like agriculture) as well as between content written on general entrepreneurship topics and green circular economics topics. The articles' contents were largely found to be positive and anticipatory.

The majority of recommendations from this study span calls to further research and analysis with the AWS suite of software as well as increased data gathering efforts. By and large, the strongest recommendation echoed here is that the iLab should focus on leveraging the keywords, sentiments and topis identified in this study to conduct A-B testing between content produced according to tacit knowledge (the current standard) and explicit knowledge stemming from this study's findings. Lastly, a call to preserve the open-ended nature of this study is strongly made so that future CEU students and public domain data analysts can engage with the challenges posed by this text and image analysis.

Introduction & Business Need

As any other start-up incubator, the iLab thrives from word-of-mouth, digital and social media exposure. These channels allow them to reach potential teams, founders, leverage ideas and identify funding opportunities for their incubatees. Alongside CEU's transition to Vienna comes the iLab's need to set up shop, offices, co-working spaces, and prominence within an entirely new environment. The Vienna transition is the broader context, while the need to understand the digital environment, online exposure and media patterns of the Viennese start-up ecosystem are some of the many problems associated with it. As part of the larger whole, this open-ended project aims to leverage web scraping, image, and text analytics to identify patterns of association in the iLab competitors' web content. The niche of this project is given by the client's interest in (but lack of awareness of) web content produced by its competitors. The guiding research question of this project is: *Who are the Viennese/Austrian start-up incubator market's competitors and what are the key digital (textual and visual) content metrics CEU's iLab should consider and benchmark when entering the new market?* Without an explicit understanding of the type of digital content created, speech used, and persuasive language leveraged by their competitors, the iLab cannot start creating industry-competing or ecosystem-challenging digital content. This intended textual analysis merely tackles one of the problems in the broader competitor analysis framework.

Digital exposure and metrics here refer to key words in articles, entity prominence in media, the overall sentiment of online mentions, the length of posts, insights on languages of communication and many more elements that can be extracted from unstructured text and visual content. The competitors are herein understood as institutions (incubators and VCs) with a keen geographic emphasis on, but not exclusive to, the Vienna-area. The desired objective is to analyze as many articles, posts, media descriptions and website pages as possible with sentiment, entity, and image analysis tools in order to identify competitor content patterns. The media concerned by this analysis spans only own-curated blog posts, news, and sponsored content. As part of further research and the continuation of this project, content about the Viennese start-up incubators could be scraped from independent media as well to compare to own-curated content. However, this endeavor is beyond the scope of the current study.

Furthermore, the scope of this project does not include prediction or trend analysis. This is also the reason behind this project's partial solution to the wider problem, that of a competitor analysis. While text analysis methodologies help uncover semantic patterns, word associations and important entities, they do not constitute any basis for predictive or prescriptive analytics

without SEO metrics. The goal of the text, token and image analysis spans topic identification, sentiment distribution and key statistically significant differences across the board of the competitors' proprietary content.

To supplement the sentiment and entity identification scope, this project takes on image analysis by employing AWS software on the visual content distributed with each blog post or news article. The findings from the text and image entity analysis are used in parallel to validate the presence of common elements and entities (such as experts, finance-talk, start-up teams, leadership, etc.) within the iLab competitors' proprietary content.

Relation to Client's Strategy & Progress

In order to tackle the new market, compete with existing incubators and attract Austrian start-ups, the iLab would need to create digital content to attract incubatees and gain prominence. The project is significant for the client in their expansion phase as it paves the way for tailoring their content creation and SEO efforts to the Austrian market. The iLab is not currently formally aware of its foreign competitors as well as Vienna's start-up ecosystem's media. Relying entirely on tacit knowledge, while having worked on a small scale for the iLab until now, is not a feasible way forward in terms of web content production. This lack of explicit knowledge and quantitative measures is what this project aims to fill, by creating reproducible and explicit documentation of keywords, sentiments, entities, topics, textual patterns, and visual content patterns. Some of the efforts that have already been made in identifying start-up incubators, VCs and summit events in Austria make up a basis of websites to start scraping.

Project Stakeholders

- *Mark Kis* – the iLab's data analyst who provided data analysis expertise and visualization creation guidelines when targeting the management team.
- *Nora Varady-Wagner* – the iLab's project manager and my direct contact with whom I had catch-up meetings to discuss interim goals.
- *Dr. Andrea Kozma* – the iLab's director, in charge of the expansion to Vienna.
- *Miklos Koren* – CEU faculty supervisor on data structure and research matters.
- *Eduardo Arino de la Rubia* – CEU text analytics advisor.
- My fellow iLab interns – *Fanni, Ashraf* and *Kerim* - who have helped put together a list of Austrian start-up incubators, VCs, and summit events that this project relied on.

Operative Goals & Expectations

The answers to the following questions, sub questions and cues for research represent the client's operative goals of this capstone, and the expectations that need to be covered in the broader competitor analysis. These points facilitated in drawing up the methodology and project plan.

General Guidance Questions (for current analysis & further research)

- Who are the Austrian market's competitors and what are their key digital metrics that CEU's iLab should consider and benchmark when entering the new market? (Facilitated by client's tacit knowledge)
    - Who are the main institutional players?
    - What kind of industries do their portfolios focus on?
- Does the current start-up market size and distribution match the start-up trends?
    - Can we model Austria's start-ups on the same distribution as Hungary's start-ups (example: dominated by FinTech or by nutrition)?
- Which are the big incubators, start-up news sources, umbrella organizations and industry references of the Austrian start-up ecosystem?
    - Can aggregate websites, portfolio websites and review websites be leveraged?

Questions Specific to the Text & Image Analyses

- What are the most popular words, expressions, and token associations in the iLab's competitors' digital content?
- Which words and expressions are highly correlated with each other?
    - What about token correlations between competitors' content?
- What does the network word associations look like in the competitors' content?
    - Use of bigrams and word-occurrences to identify key expressions.
- What is the prevailing sentiment across media articles about the start-up ecosystem?
    - What are the most sentiment-loaded words?
    - Which incubator uses the most sentiment-loaded language?
- What topics can be modelled from the pool of competitor content?
    - Employ topic modelling to find "naturally-aligning" clusters of content.
- What are the main languages of communication?
    - What does the English to German distribution look like within the articles?
    - AWS comprehension algorithms to benchmark sentiments across languages.

- What are the main entities identifiable within visual content created and distributed by start-up incubators in Austria?
    - Identify main visual elements through AWS machine learning.

Methodology & Strategy

The following section outlines the data gathering, cleaning and analysis' explicit assumptions and processes used throughout the study.

Open-Endedness

This study is open-ended, meaning that there are no formal hypotheses to be made with regards to the findings. The goal is that of providing descriptive analytics. Accordingly, the study's exploratory data analysis (EDA) nature means that it should help the client identify potential patterns of association to inspect further and/or a lack thereof.

Representative Sample Assumption

The base data needed for content gathering and analysis spans competitor identification in the form of shortlisting incubators, news aggregators and start-up events concerning the Austrian market. This identification was conducted through discussion with the iLab, the client having expressed a high interest for all sizes of incubators, through qualitative research (reading articles, following references, word-of-mouth etc.), and through work with the iLab interns who put together a list of most of the major Austrian start-up accelerators. Furthermore, the list of official start-up accelerators put together by the Vienna Business Agency supplements the list of incubators used in this study. Finally, the list is formatted under a tacit classification of whether the company provides digital content in the form of a blog, news, or events page, separating by language of communication.

Start-up Incubator	**Content Language(s)**	**Own Content**
INiTS	English & German	☑
A1	English & German	☑
Agro Innovation Lab	English & German	☑
Blue-Minds Company	German	☑
Greenstart	German	☑
ImpactHub	English & German	☑
TechHouse	English & German	☑
Factory1	English	☑
i2c	English	☑

InvestmentReadyProgram	English	☑
MatchMaker Ventures	English	☑
Tech2Impact	English	☑
The Ventury	English	☑
Austrian Startups	English & German	☑
sic!	N/A	☒
I2A	N/A	☒
Epiphanic	N/A	☒
i5invest	N/A	☒
HK Incube	N/A	☒
Elevator Lab	N/A	☒
weXelerate	N/A	☒
INNO VATE	N/A	☒

Table 1

This study only focuses on the first 14 incubators, as they produce their own online content under the form of blogs or news. The rest of the incubators are beyond the scope of the analysis and, thus, do not make up the sample for gathering through web scraping. The term 'competitors' is hereinafter taken to mean only the 14 incubators listed above that produce their own content.

Data Gathering

The data comes as unstructured scraped texts from competitors' websites, blogs, and other outreach channels. The features of interest are the article links, image links, article titles, article contents and publishing dates. The web scraping scripts serve the purpose of gathering all this data and storing it in CSV files for each incubator. Once all the designated incubators have been scraped, the content CSVs are concatenated into a large data frame (each row having a unique ID) on which data augmentation, processing and analysis can be done. The image links are used to download the images associated with every scraped article and store them in an artefacts folder with the unique IDs as names.

Data Cleaning & Augmentation

The large data frame of all competitors' content is supplemented with language recognition, language classification accuracy and sentiment score columns through data pre-processing scripts leveraging AWS Comprehend. These columns serve analytics purposes as they provide quantitative measures like rate of positive words and likelihood of language usage.

Pre-Pilot

With the tacit work paving the way for the technical analysis, a web scraping strategy was devised with Miklos Koren, the faculty supervisor, to run a pre-pilot on a smaller portion of data gathered from a (representative) sample of the competition. This pre-pilot is used as the basis for the bulk of the text and image analysis. While the sample of articles chosen represents an arbitrary decision, the pre-pilot only serves technical purposes of developing the scalable text analysis scripts. Thus, any bias associated with the data cherry picking is waived once the pilot analysis is be done on the full sample (arguably even population, as this study scraped all the news and blog contents of each incubator with the use of scalable scripts).

Analysis

On the AWS image entity analysis level, an AWS Rekognition script analyzes the photos stored on AWS S3 and joins this entity dataset back to the large dataset of all articles to conduct analytics and visualizations. On the AWS text analysis level, the augmented dataset containing language and sentiment columns is used for analytics and visualizations. On the R text analysis level, the general-purpose analysis lexicons available within the R sentiments package are supplemented by an open-source sentiment lexicon for the German language. The stop-words lexicons are also borrowed from open-source packages. The bulk of the data used in this project spans textual tokens all converted into tidy format for processing and analysis.

Validation

The result validation is three-fold. First, the process validation occurred under the form of progress meetings and emails with the client and advisor to discuss interim goals and progress. Naturally, this type of validation is largely implicit, as it relies on the client and advisor's understanding of the Austrian start-up ecosystem.

Second, an internal validation of the text analysis models was conducted by employing AWS services such as Comprehend. This part of model validation was not discussed with the client as it is more important for token analysis fact-checking (i.e., benchmarking results). The black-box nature of the AWS services is the reason why this project heavily relies on in-house text analysis scripts (i.e., transparency and reproducibility for the client).

Lastly, the third type of validation concerns the external validation which is far beyond the scope of this limited project. A strategy has been devised with Eduardo Arino de la Rubia and Miklos Koren. The most common words, expressions, topics, associations, and entities stemming from the analysis can be evaluated externally by doing A-B testing using SEO tools on content produced in-house by the iLab. As such, the iLab can create content to test on the

Austrian market, some using this analysis' identified common phrases and some relying entirely on tacit knowledge for content creation. The success of both types of content as well as keyword prominence can be measured later with SEO tools. However, this external validation is beyond the scope of this capstone project.

Repositories & Version Control

Git was chosen for version control and data syncing. Two GitHub repositories were created, one for data gathering and pre-processing and one for data analysis and visualization:

- Scraping Austrian Incubators – bulk of data, images, scripts, and HTML artefacts,
- Analyzing Austrian Incubators' Content – analysis scripts and visualization artefacts.

Tools

The programmatic tools used in this project are:

- R for web scraping, data gathering, cleaning, statistical programming, text analysis:
  - *dplyr, data.table, tidytext* and *tidyverse* approaches,
  - *revest* and *httr* web scraping packages,
  - *ggplot, kable* and *ggraph* visualization packages,
  - *stringr* for regular expressions and string manipulation.
- Python for automated webpage download:
  - *selenium* package used in concurrence with *Google Chrome driver*.
- AWS S3 for scalable file storage:
  - *aws.s3* R package with AWS-issued access key.
- AWS Rekognition for image entity analysis:
  - p*aws.machine.learning* R package with AWS-issued access key.
- AWS Comprehend for language recognition and sentiment analysis:
  - *aws.comprehend* R package with AWS-issued access key.

Deliverables

The main deliverable of this project concerns this written report which contains a detailed description of the methodologies used, the results, visualizations created based on those results and recommendations for further course of action. The final presentation covers the same structure as the technical report. The final deliverable is a project summary to be officially submitted and stored in CEU's repository of theses for archiving and administrative purposes.

Project Reproducibility

The analysis is covered in-depth in the final technical discussion deliverable of the project, describing all the assumptions made and steps taken within data gathering and analysis. The iLab team also received the web scraping scripts written as functions that are easy to re-run on their local machines by directly calling on the source codes. Furthermore, all the code is pushed to a GitHub repository shared between all the stakeholders.

Project Plan

The following stages represent the technical processes required to obtain answers to the research question(s):

1. Reproducible and scalable web scraping scripts need to be created for the bulk of the competitor websites.
    a. Built to gather both the text and the image data from each article.
    b. Create CSVs of all scraped content.
2. Bind, clean and augment data with AWS Comprehend features.
    a. Create CSVs of data at each stage.
3. AWS text analysis and image analysis on the full sample.
    a. Cross-language comparison.
    b. Cross-sentiment comparison.
    c. Cross-incubator comparison.
    d. Create visualizations in artefact repository.
4. Pre-pilot text analysis needs to be conducted on the arbitrarily designated sample (not included in "Analytical Steps" section below).
    a. Sentiment analysis, topic modeling, tf-idf analysis and n-gram associations are done with the proprietary R *tidytext* package.
    b. Result validation and benchmarking happens with AWS Comprehend for sentiment analysis.
5. Text mining and image analysis scripts need to be made into scalable format to be run on the full sample.
    a. For the German language texts, the token analysis scripts need to be made in a scalable fashion so that they allow the interchanging of stop-words and sentiment datasets.
6. The full-sample text analysis consists of:
    a. Token analysis,

b. Word correlation analysis,
c. TF-IDF analysis,
d. Cross-topic analysis,
e. Cross-competitor analysis,
f. N-gram analysis.
g. Cross-sentiment comparisons.
h. Create visualizations in artefact repository.

7. Create documentation; contains the project's technical discussion.

Data Gathering & Manipulation

The following section concerns the technical discussion on data gathering and manipulation. The affiliated GitHub repository is the 'Scraping Austrian Incubators'.

Web Scraping

The R *rvest* package was mainly used for scraping. In total, there are 14 R scraping scripts in the GitHub repository, each tailoring to the website structure and HTML peculiarities of each incubator. The following script can be inspected to follow this technical discussion on scraping. Two functions were defined for scraping, one to gather the article and image links for every article displayed on every news page (for the incubators with very old content, 2016 was picked as the earliest year for an article's publishing date) and one to extract the title, content body and date of each article. The HTML containers and objects that the elements of interest were nested within were called inside of the *rvest* functions. Regular expressions were used to clean up the residual text.

Each incubator's two scraping scripts in turn created one CSV each and placed them in the incubator's respective file folder within the GitHub repository (file folder for example script). The following tables summarize the features of each CSV table.

Feature	Example	Type
Article_link	"https://vienna.impacthub..."	String
Img_link	"https://vienna.impacthub..." OR "no_image"	String

Table 2: Features of article_links.csv template table

Feature	Example	Type
Article_link	"https://vienna.impacthub..."	String
Creator	ImpactHub	String
Article_title	qualitätszeit gegen die digitale Kluft	String

Date	January 1, 2021	String
Content	"A mad man lived in the 19th century of …"	String
Img_link	"https://vienna.impacthub..." OR "no_image"	String

Table 3: Features of articles_content.csv template table

Selenium Addendum

For some incubators' websites, scraping directly with R was not possible due to infinite scrolling features and other Javascript elements on top of the standard HTML objects. This was the case for the websites of The Ventury, Tech2Impact and Blue-Minds Company. The following script can be inspected to follow this technical discussion on scraping. A Python Selenium function was written to automate a Chrome WebDriver to scroll all the way down to the bottom of the news pages and then download the HTML (all of this was done headlessly for scalability and replicability's sake). The R scraping scripts outlined above were then applied on the downloaded HTML files.

Merging the Datasets

All the *articles_content.csv* files were iteratively read and bonded together to produce the main data frame of this study. A unique ID column was added for each article. All the augmentations and analyses were done on this base data frame.

Availability and Quality of Data

To control for formatting issues, the UTF-8 encoding standard was used, white spaces were removed, unknown characters were removed as well as all repetitive promotional lines of content. Nonetheless, the quality of the data is about the standard of what would be expected from unstructured raw text written in many different formats. One downside to this web scraping effort was that the values produced for the date column were not standardized, thus rendering the date column hard to use for time-series analysis out-of-the-box. Nonetheless, that was beyond the scope of this study.

Data Augmentation with AWS

The base data frame of incubators' content was first augmented with the addition of two columns pertaining to language. AWS Comprehend's feature to detect language was used by connecting to the AWS services with the provided public and private access keys. The following script detects the language of the articles and binds the *LanguageCode* and *Score* columns to the main data frame. The latter column contains likelihood values in percentage for the identified language.

The second augmentation was done with the help of AWS Comprehend's feature to detect sentiment. The following script identifies the sentiment percentage of each article's content. Because of AWS' 5000-character limit per API call, a function was devised to split the longer articles into chunks of at most 5000 characters, analyze their sentiment and then combine the values back on a per-article basis. The AWS sentiment script binds the *negative, positive, mixed,* and *neutral* columns to the base data frame. Each of these columns represents a percentage of the total sentiment, thus the values for these four columns sum up to 1 (i.e., 100%) for every row.

With these two augmentations, the augmented base data frame has the following structure:

Feature	Example	Type
ID	1	Number (not important)
Article_link	"https://vienna.impacthub..."	String
Creator	ImpactHub	String
Article_title	qualitätszeit gegen die digitale Kluft	String
Date	January 1, 2021	String
Content	"A mad man lived in the 19th century of …"	String
Img_link	"https://vienna.impacthub..." OR "no_image"	String
LanguageCode	en	String
Score	0.99	Number
Mixed	0.99	Number
Neutral	0.99	Number
Negative	0.99	Number
Positive	0.99	Number

Table 4: Features of incubators_raw_content_languages_sentiment.csv template table

Lastly, a new data frame was created using AWS Rekognition's feature to detect image entities. The following script downloads all the incubators' photos, uploaded them to AWS S3 (mandatory part of the process), runs entity recognition image by image and binds the results back into a local data frame. The table produced takes the following form:

Feature	Example	Type
Name	Plane	String
Creator	ImpactHub	String
Confidence	0.99	Number
ID	1	Number (not important)

Table 5: Features of incubators_images_10_entities.csv template table

The data gathering has thus produced one data frame where each observation is an individual article and contains language and sentiment values, and one data frame where each observation is one entity identified in a certain article. The data analysis part builds solely on these two data frames.

Data Analysis

The analysis section is divided between the AWS-based analyses and the R-based analyses. The first part, content text analysis, relies on the language and sentiment columns previously augmented to the base data frame. The second part relies on the image entity data frame produced by the AWS Rekognition function. The last and largest part relies on the augmented base data frame to run token, n-gram, topic, correlation, and sentiment analysis using R's *tidytext* approach. The following repository contains the analysis scripts.

Content Text Analysis with AWS

The AWS content analysis can be followed in this script.

Character & Article Distributions

First off, it is worthwhile to look at how many articles we have for every start-up incubator within our sample of 819 articles. Since the scraping was done on the full content of each incubator's news page, it is a fair assumption to make that these numbers of representative of all the articles each incubator wrote.

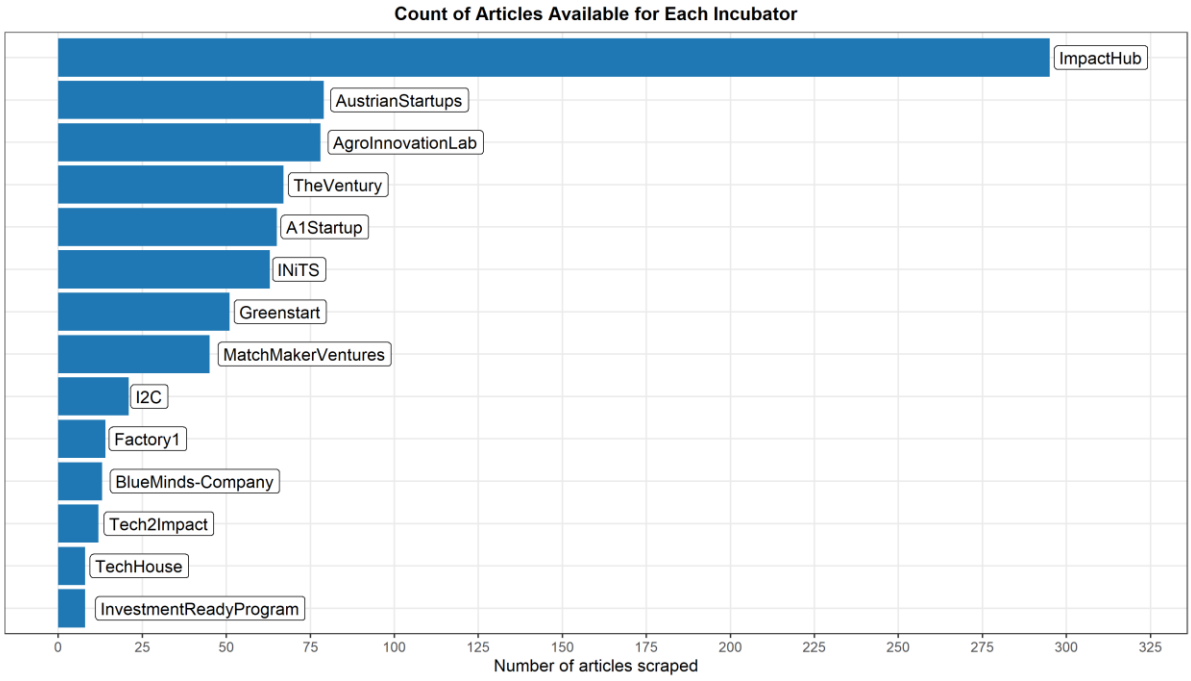

Figure 1

Figure 1 shows that the distribution of articles by incubator is largely skewed. Impact Hub has almost 300 articles written while IRP and Tech House each have a mere 8 articles written. On the topic of distributions, it is important to look at the histogram of character counts.

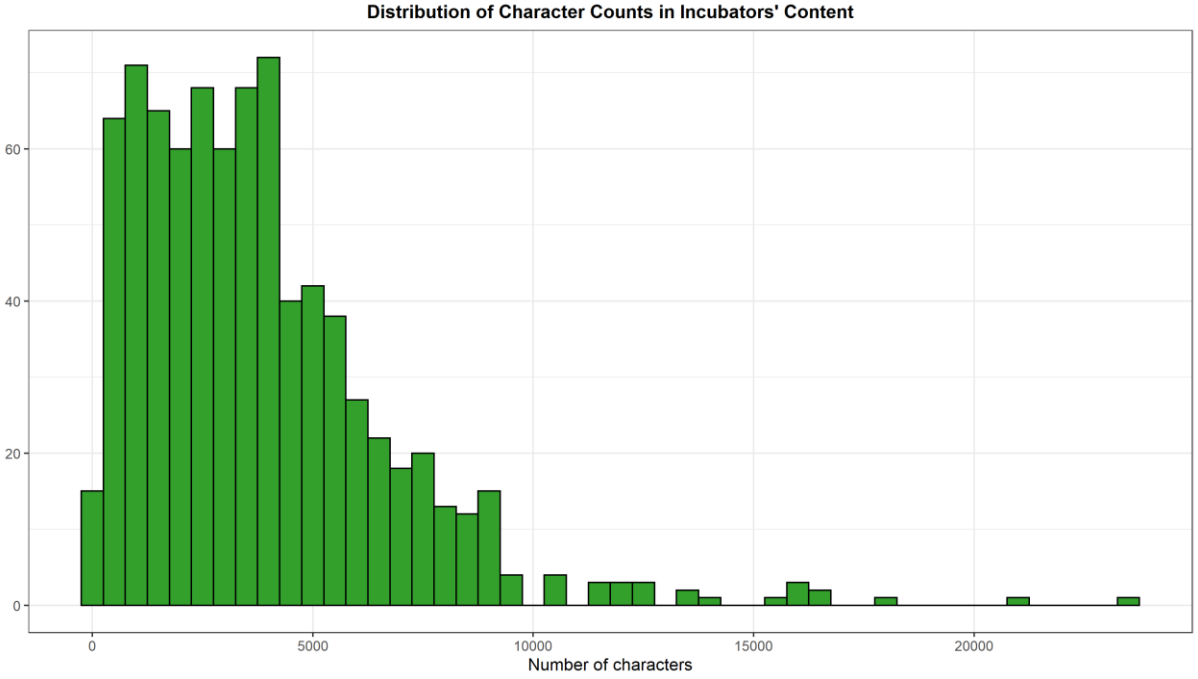

Figure 2

Figure 2 shows a normal distribution of characters, with a few articles having excessively long bodies of content. While the character count distribution is useful to gauge whether there are many outliers, a deeper dive into word counts per incubator provides more insight into the which incubators have extremely long articles.

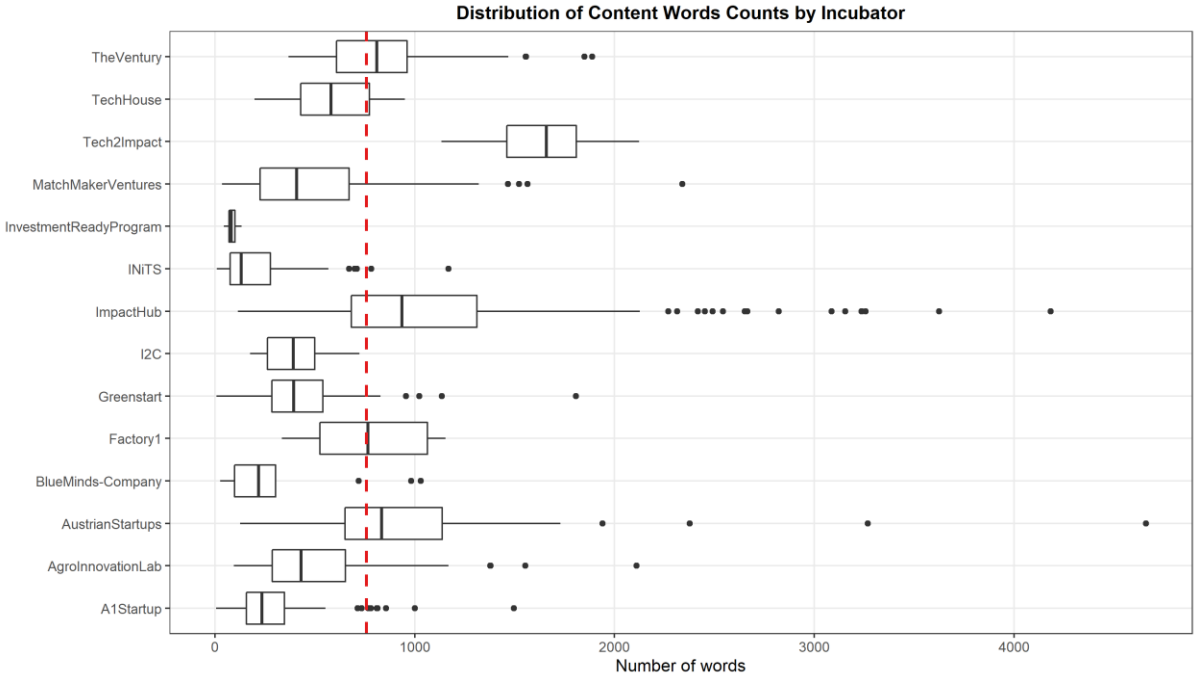

Figure 3

With a mean length of 850 words, most articles are rather short. The incubators with the most articles also seem to have the most uncommonly large articles. In fact, it is only the Investment Ready Program that has extremely short articles, only a few sentences, and is dragging the mean down. The rest of the incubators seem to agree on an average article length between 500 and 1000 words (or an equivalent of 5 paragraphs).

Language EDA

First off for language analysis, one can inspect the count of articles by language per incubator.

Incubator	**German Articles**	**English Articles**	**Total Articles**
A1 Startup	64	1	*65*
Agro Innovation Lab	45	33	*78*
Austrian Startups	9	70	*79*
Blue-Minds Company	13	0	*13*
Greenstart	51	0	*51*
Impact Hub	29	266	*295*
INiTS	46	17	*51*
Tech House	4	4	*8*
Factory1	0	14	*14*
i2c	0	21	*21*
Investment Ready Program	0	8	*8*
MatchMaker Ventures	0	45	*45*
Tech2Impact	0	12	*12*
The Ventury	0	67	*67*
Overall	*261*	*558*	*819*

Table 6: Articles for each language by incubator

Articles written in English clearly dominate the distribution, irrespective of Impact Hub's huge contributions to both languages. For the content text analysis, the individual language columns were created for English and German to compare percentage differences.

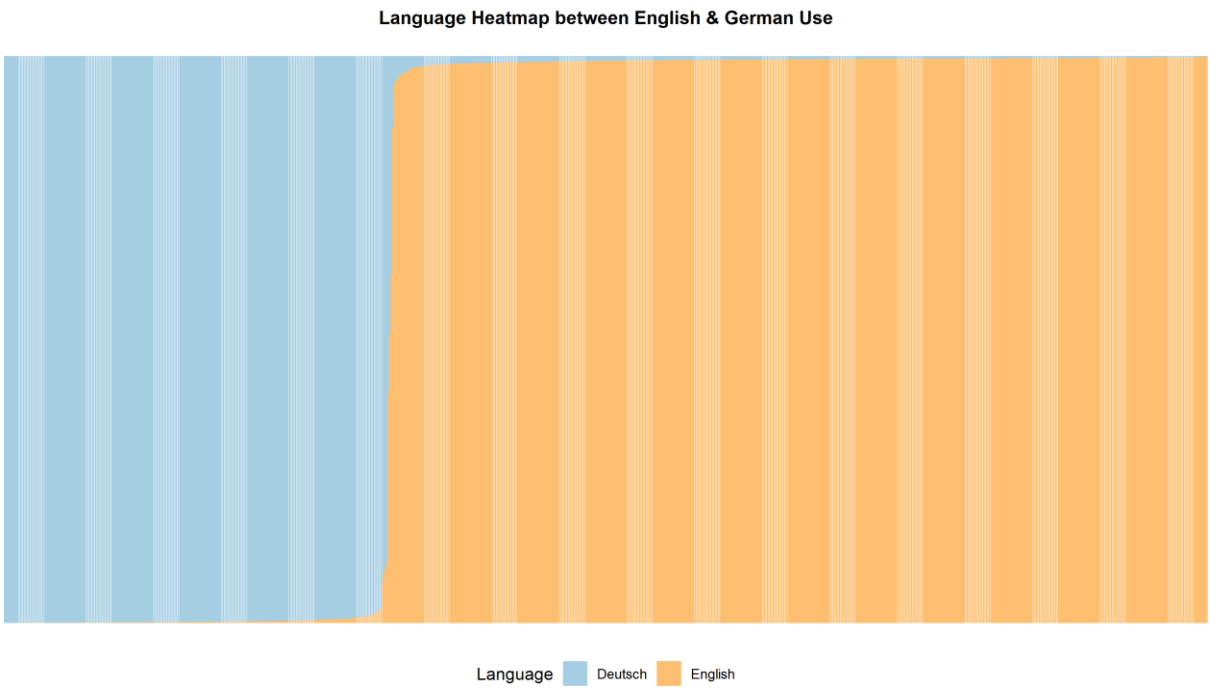

Figure 4

The heatmap above shows a very homogenous distribution between the percentage likelihood of English and that of German. This also begs for the explanation that the AWS machine learning algorithms do not provide a word-by-word classification, but rather an advanced black-box language classification.

As such, mixed language articles are far and few in between in this representative sample of Austrian start-up incubators' content. It is curious to see if this lack of language-mixing is present in the article titles as well.

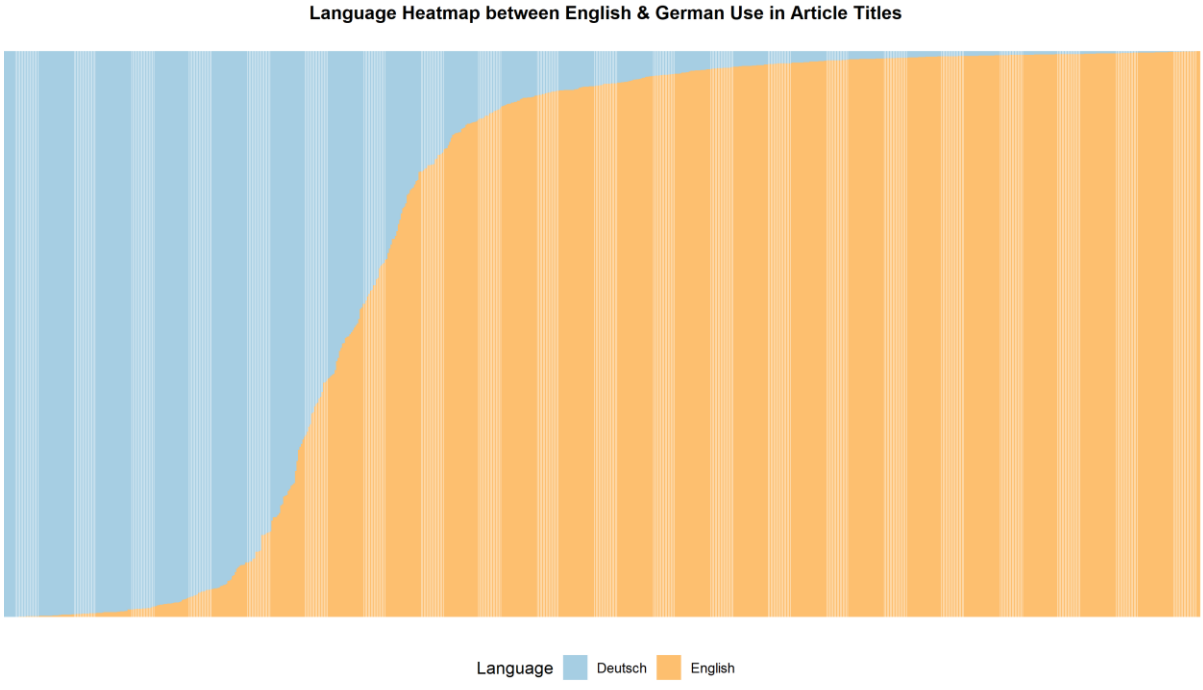

Figure 5

The language mixing tendencies show up much more in the articles' titles, potentially meaning that start-up incubators like to use common English expressions and known words in their article titles even when the article itself is in German. The opposite could be true for German, but it is less believable.

Sentiment Analysis

On the topic of sentiment analysis, the most important visualization is the distribution of articles by the percentage of each sentiment.

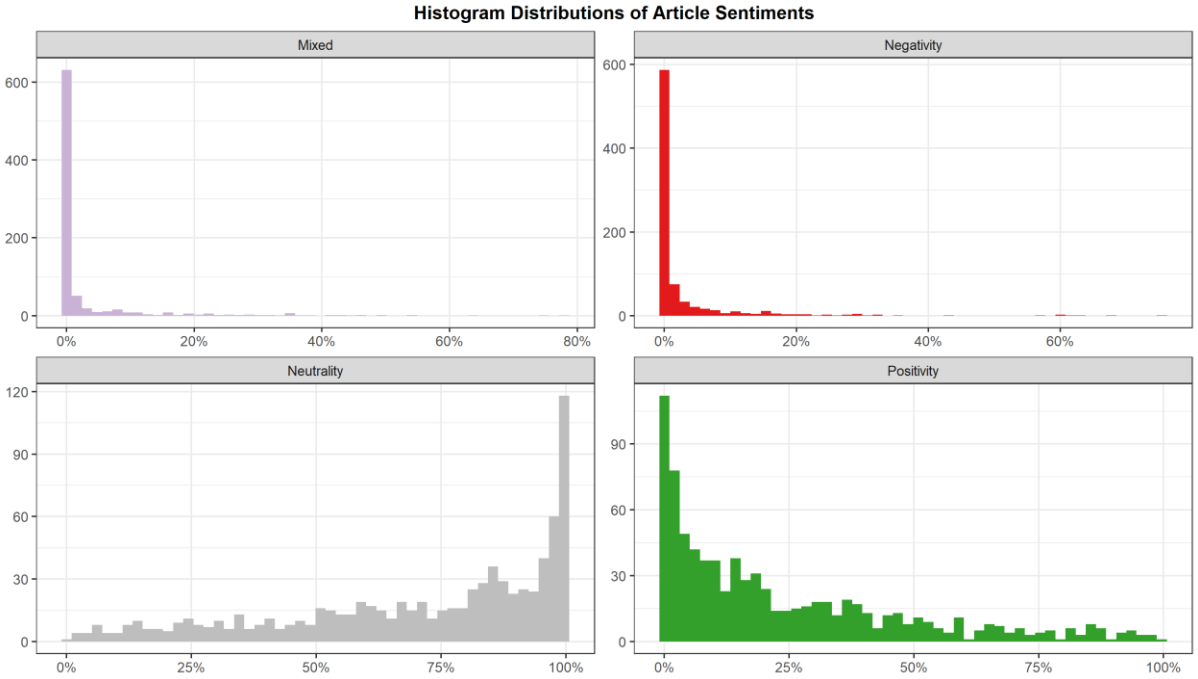

Figure 6

The majority of articles get a neutral score, while a fair amount seems to have a significant proportion of positivity. However, visualizing these differences on a stacked scale would better showcase what proportion that negative and positive sentiments take on average.

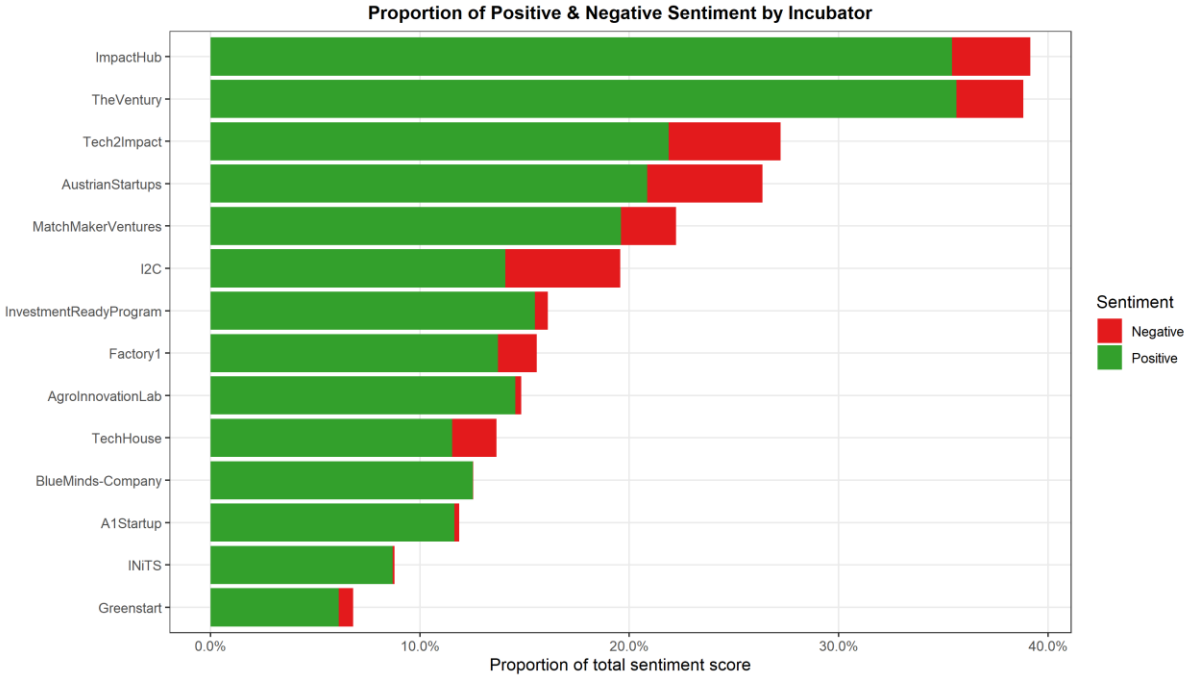

Figure 7

Figure 7 shows that incubators like Impact Hub and The Ventury have an average of 35% positivity and around 5% negativity in their articles, thus proving to be the most sentiment-loaded incubators by content production. The more specialized incubators such as Greenstart and Agro Innovation Lab seem to have much less prominent sentiments in their articles, perhaps using a lot more neutral language. Knowing that Greenstart and Agro Innovation also have articles in German does raise the question of whether articles written in English were found to be more sentiment-loaded.

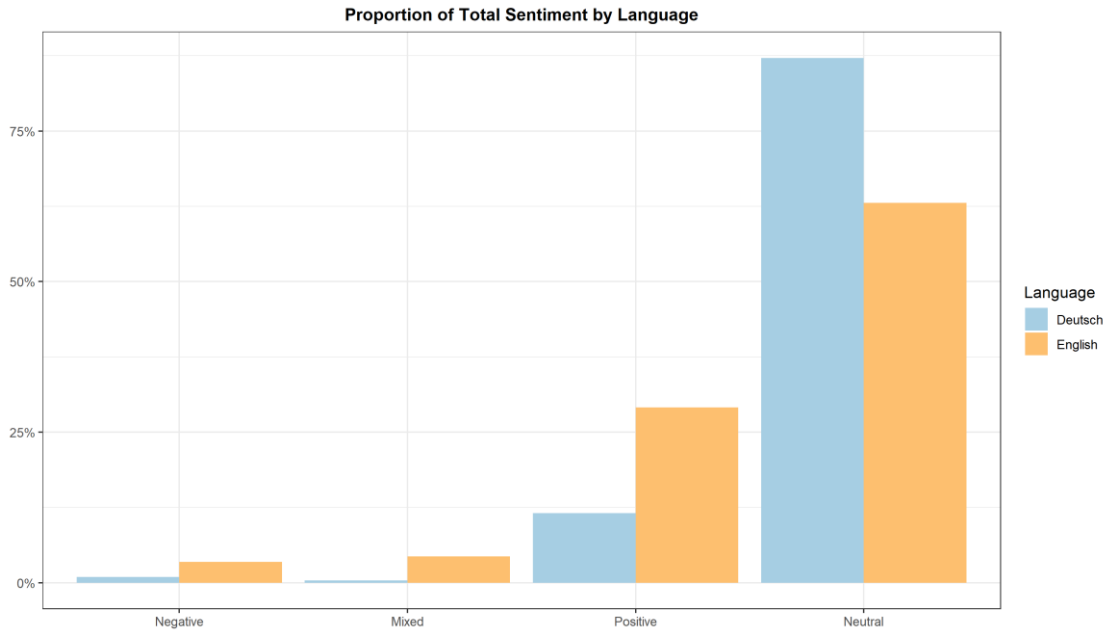

Figure 8

Figure 8 seems to showcase the hunch from before, namely that the German articles are a lot more neutral on average. Taking a deeper look plotting all articles on a scatterplot:

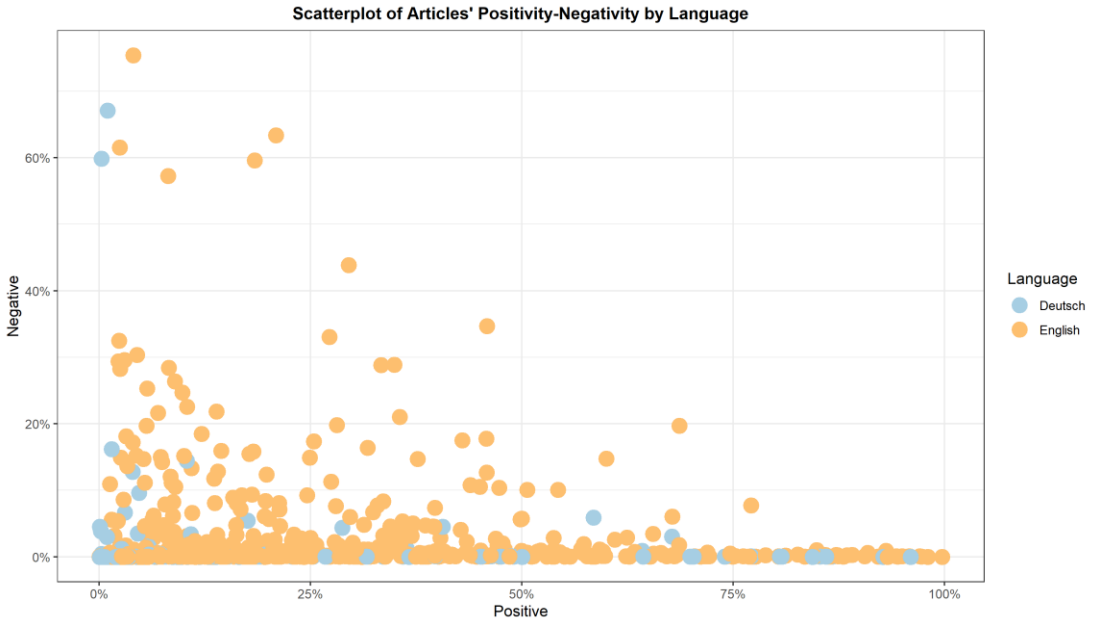

Figure 9

It seems like there are only two articles in German in this representative sample of start-up incubators' content with a negativity score over 20%, while there are a lot more in English. Overall, English seems to be identified as much more often mixing the two emotions, while German seems a lot more one-sided. Following in the scatterplot trend, plotting the axes but this time each point being an incubator should give a clearer idea of where each incubator is situated on the sentiment scale.

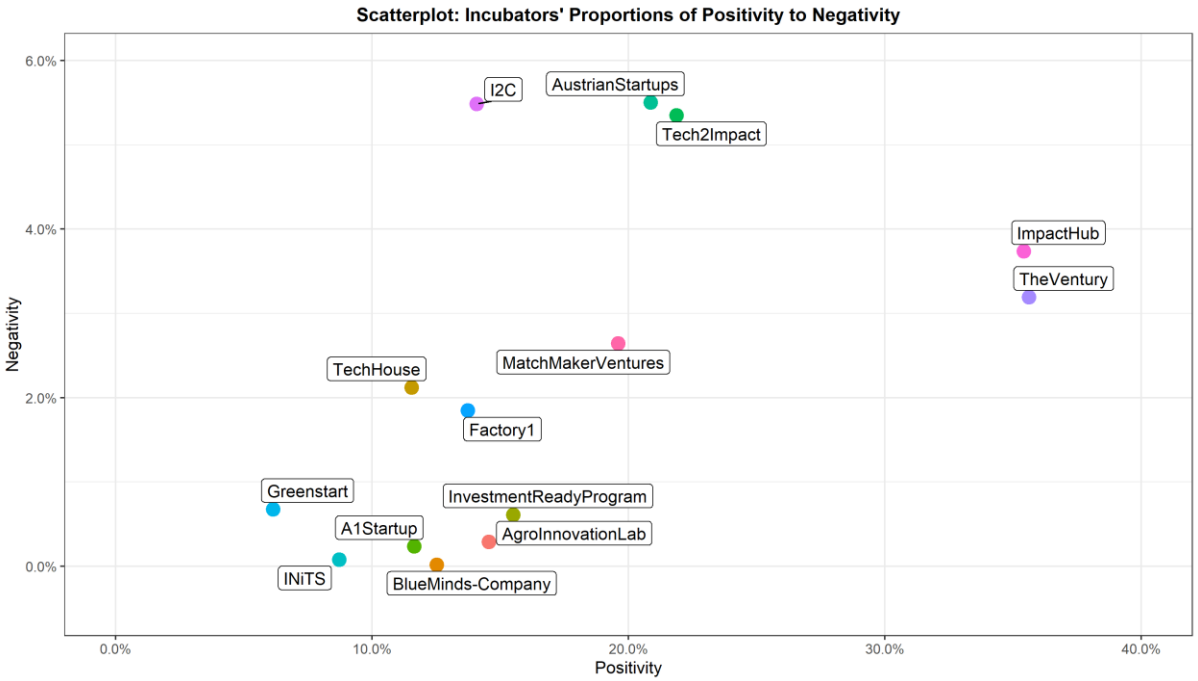

Figure 10

The same graph but with neutrality on the y-axis can be found in the appendix. Figure 9 allows us to identify three "clusters" between the incubators when comparing content sentiment. Austrian Startups, i2c and Tech2Impact seem to have their own quadrant of slightly higher negative language and less positive language overall.

Lastly, following the positivity-negativity scatterplots, it is worthwhile to identify the creator of the most sentiment-loaded articles. To find the true outliers in this case, only articles that scored over 15% on both negativity and positivity are plotted.

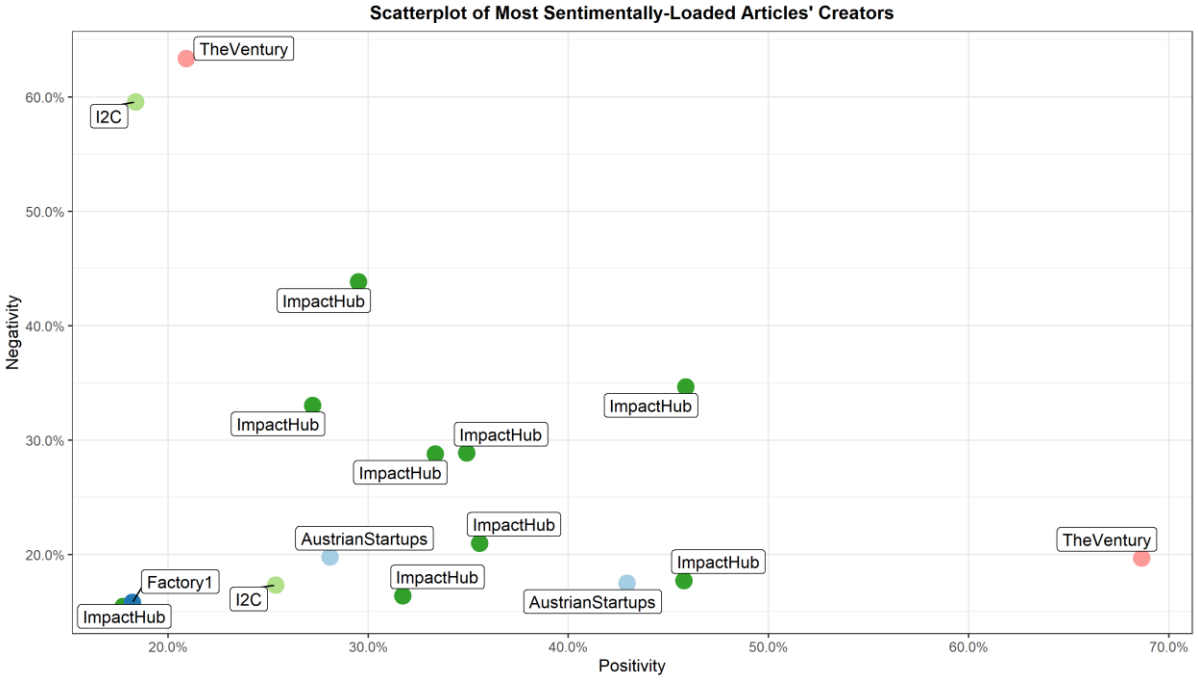

Figure 11

Overall, the takeaways from this sentiment analysis section are that incubators like The Ventury, i2c, Austrian Startups and Impact Hub produce the most sentiment-loaded articles in the Austrian start-up ecosystem. On the language spectrum, all of these incubators largely produce content in English, which was also found to exhibit more positive and negative sentiment on average in the articles. Perhaps the German articles are a lot more specialized, thus using more industry and technical terms.

Image Entity Analysis with AWS

The following section concerns the image entity analysis conducted on the augmented base data frame which was merged with the AWS Rekognition dataset. The analysis can be followed in this script. As Tech House and Investment Ready Program did not have images linked to their articles, they are not part of the analyses below.

Entities & Sentiment

On the topic of image entity analysis, it is worthwhile to combine the recognized entities with their associated article's sentiment. To compute a positivity score and a negativity score and select the top 20 image entities from each sample to compare and see if (and how many) overlaps there are. The potential drawback with this approach is that the most frequent entities might dominate the charts, even though the entities themselves might not be loaded with many emotions.

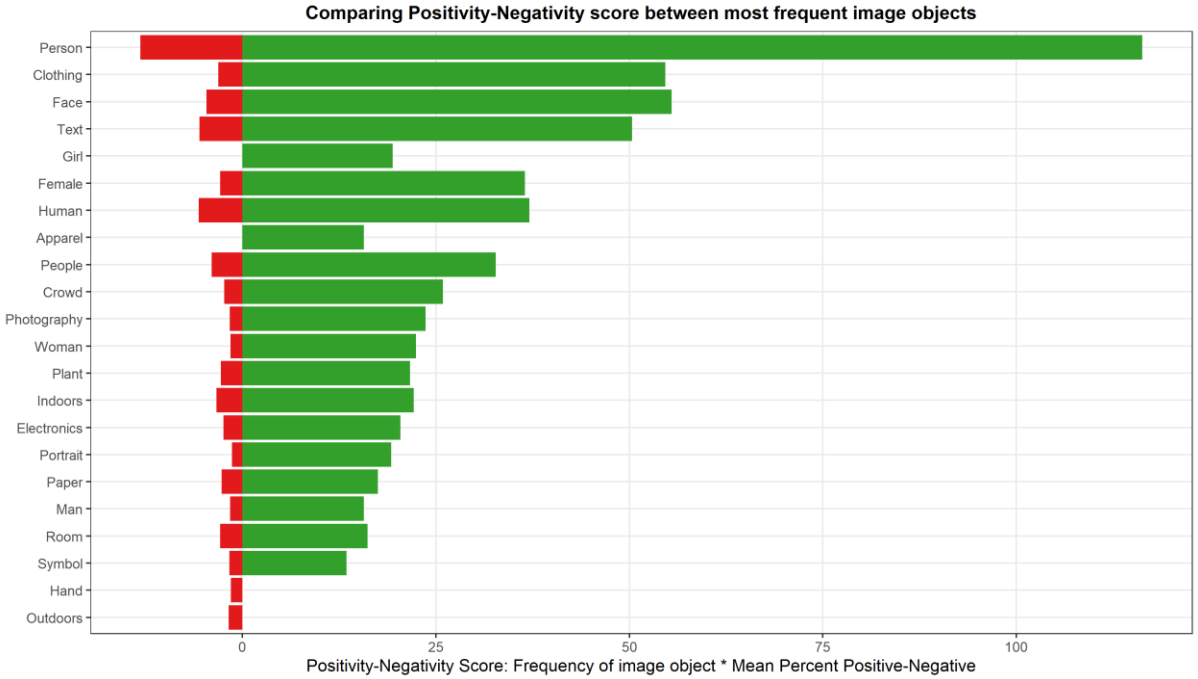

Figure 12

Figure 12 proves the hunch from before that using frequency as part of the computation means that the top entities by frequency dominate the spectrum. Naturally, elements like 'person', 'text', and 'photography' are not loaded with much sentiment if one cannot know what the person is doing, what the text writes or what the photo is being taken of. Another approach would be to directly use the mean AWS Comprehend sentiment percentage scores.

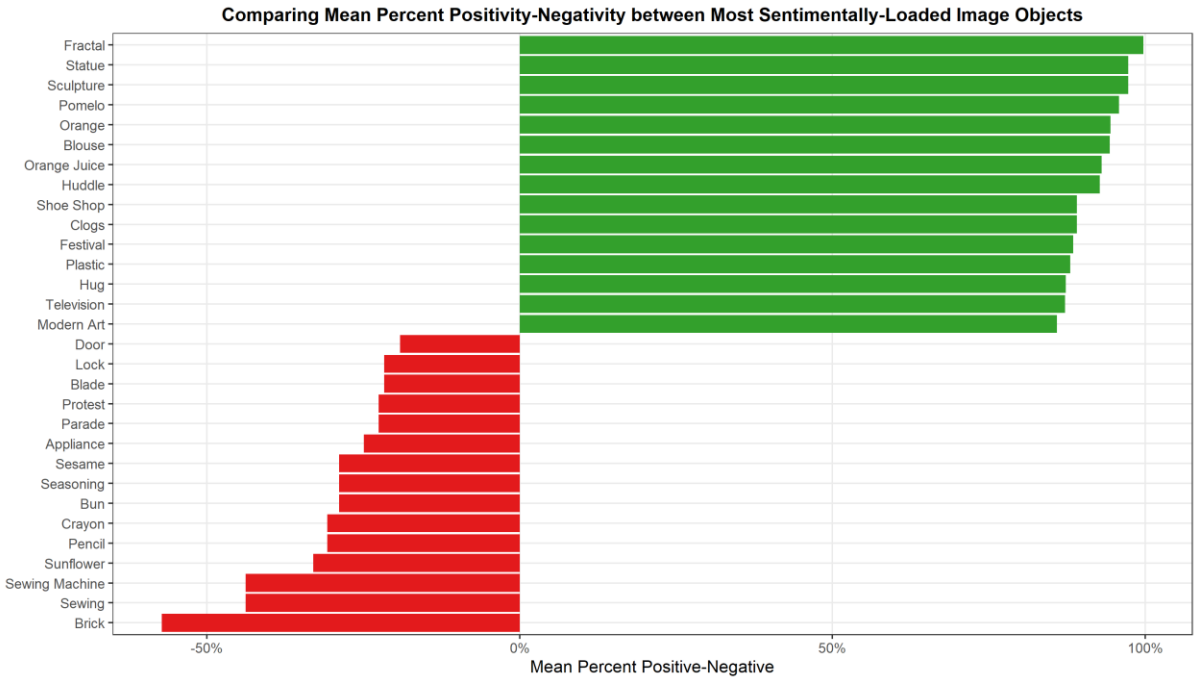

Figure 13

Using the mean percentages yields much more interesting results, with a lot of fruits (pomelo, orange, and orange juice) being associated with the most positive articles, while common household items like a sewing machine, appliances and blades seem to point towards the much more negative articles. The presence of elements like hugs for positive emotions and protests for negative emotions at least tacitly (i.e., by human feeling) validate the findings of Figure 13.

Top Entities

It is worthwhile to also inspect what the top entities are by frequency.

Figure 14: Word cloud of top image entities by count

Unfortunately, the word cloud above showcases the main drawback when using an AWS machine learning service, which is that very simple entities like people, faces, text, humans, and crowds will always be found, as they are also very common photo elements. As such, it is important to find a way to visualize the relative importance of each entity, rather than their overall importance. This is where tf-idf comes into play, short for term frequency by inverse document frequency.

Entities & TF-IDF

Zipf's law states that the frequency that a word appears is inversely proportional to its rank. As such, it makes sense to inspect the tf-idf to identify the "special" entities. This is exactly what a tf-idf analysis does, it weights the entities that show up a lot for many incubators lower than the entities that show up for fewer incubators. This tf-idf value can be plotted to find the most characteristic image entities between the incubators.

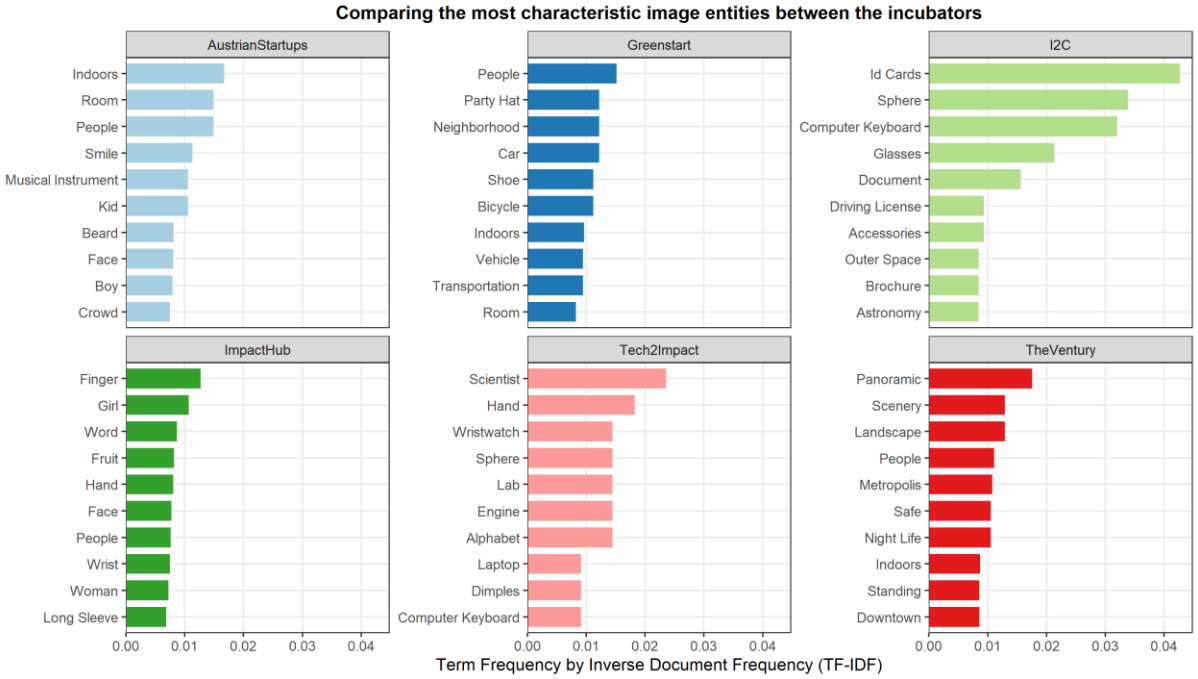

Figure 15

Figure 15 provides some useful insights into the most characteristic entities per incubator. With a potential obsession for close-up shots (of entities like hands, wristwatches, and dimples) and science-oriented entities like engines, labs, and scientists, Tech2Impact stands out from the crowd. The Ventury appears to have a different style showcasing landscapes and cityscapes quite often. Perhaps they represent the stereotypical motivational backgrounds.

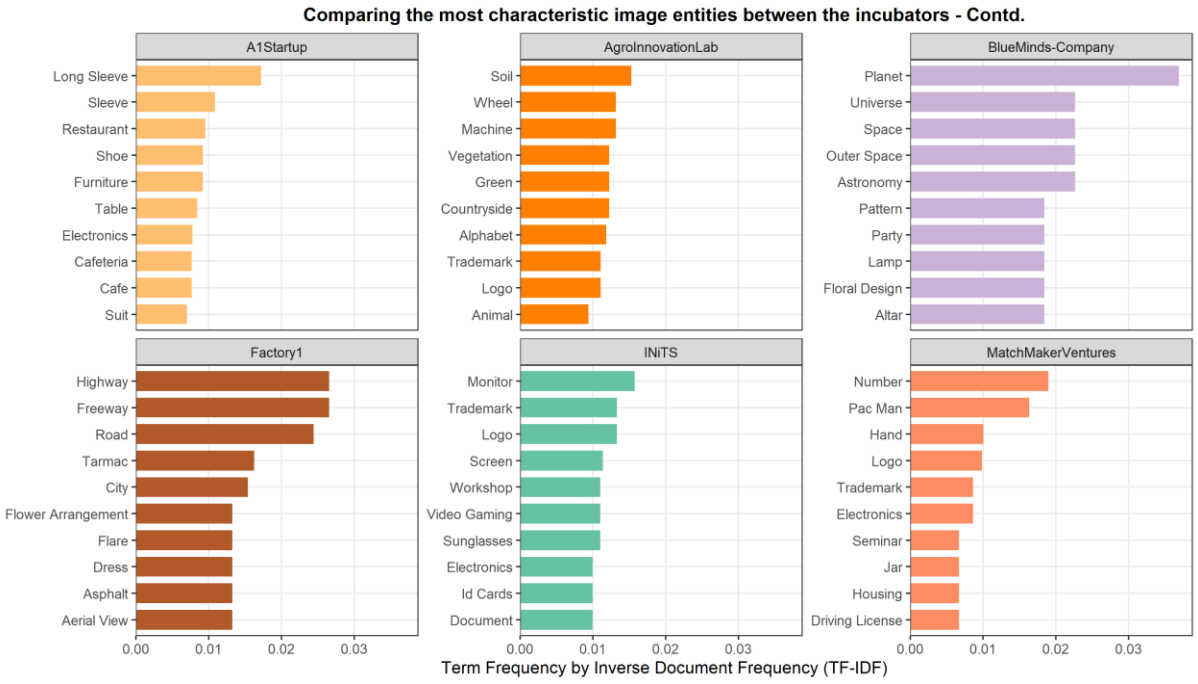

Figure 16

Continuing the plots of characteristic entities for the other six incubators, Blue-Minds Company seem to be quite fond of space-themed images while Agro Innovation Lab is staying true to their nature and sharing images of soil, agriculture, vegetation, and the countryside. With these tf-idf findings in mind, it is worthwhile to check if one can identify the specific incubator by the characteristic entities present in its distributed content.

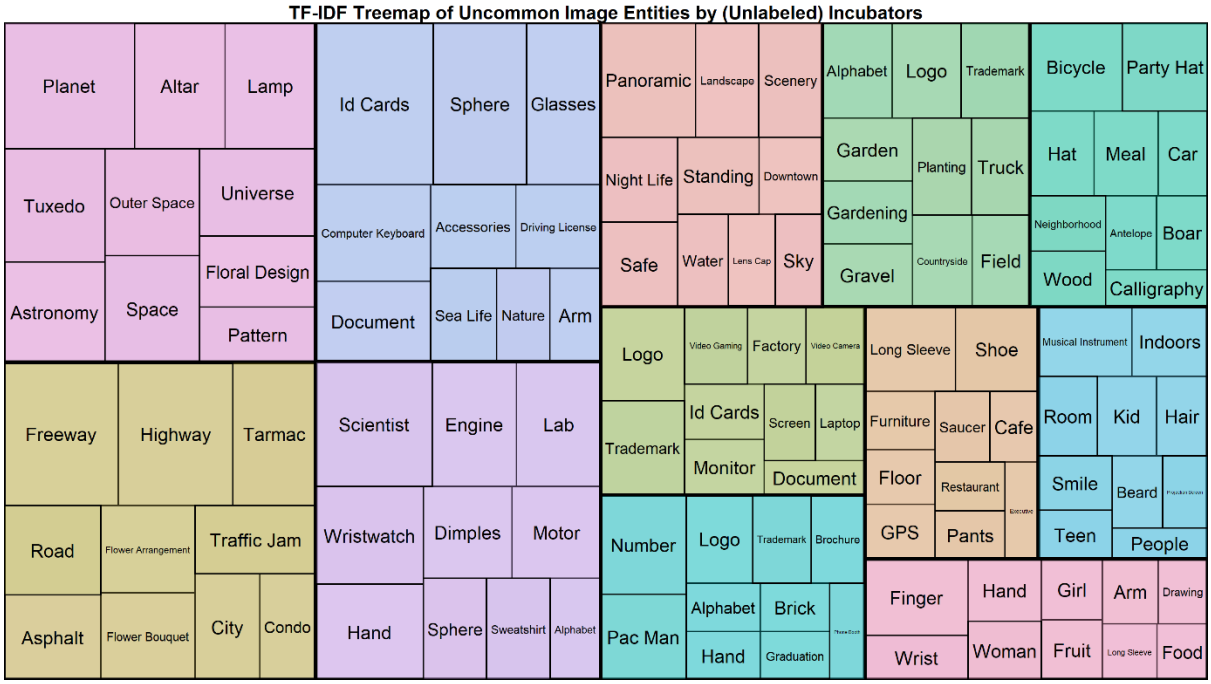

Figure 17

Figure 17 provides a slight challenge at first sight, but a closer look and a slightly formed eye on the previous findings and one is able to distinguish between some of the categories and attribute them to one of the 12 incubators. The most promising part about these AWS services is that with more scale comes better performance and identification. Ultimately, with more occurrences, the tf-idf starts to matter a lot more and even more special entities might prevail.

Token Text Analysis with R

The third section of the Austrian start-up incubators' content spans the usage of the R *tidytext* package for token analysis, word correlations, tf-idf, topic modelling, word pairs and, finally, token sentiment analysis. The following script contains all of the EDA steps taken to create the visualizations and draw the findings.

The data leveraged in this section spans the AWS sentiment and language augmented base data frame as well as open-source stop word lexicons and sentiment lexicons. For English, the open-source sentiments (*nrc, bing, afinn* and *loughran*) and *stop_words* datasets available in the *tidytext* package are employed. For German, *stopwords-iso*'s *stopwords-de* public

GitHub repository and Rachael Tatman's Kaggle sentiment database are employed. Formal references to these resources are available in the 'References' section.

The data was prepared for analysis by filtering out the instances where the content was shorter than three full sentences and splitting the sample into three groups that were each merged with the stop words lexicons concerning their language(s): *base data*, *base data EN* and *base data DE*. The *unnest_tokens* function was applied to each dataset, making them into tidy formats of one-word-per-row.

Token Analysis

It is worthwhile to evaluate the results considering the distribution of articles is highly uneven (see Table 5 above). It is important to keep this in mind when doing analyses that do not compare relative but absolute values such as token frequencies. Accordingly, the first part concerns an EDA on the most common words for both languages.

Figure 18: Word cloud of top 100 German words

Figure 19: Word cloud of top 100 English words

The top words are all highly anticipated for articles aiming to promote start-ups, incubators, funding, social entrepreneurship, etc. Not only that, but a simple ranking of all the words includes words like ‘food’, ‘company’ and ‘day’, which are all extremely common. To find most interesting results, a deeper dive is required.

Since the sample is plagued by skewness due to content creation disproportions, the analysis should look at relative frequency of words for each incubator. The goal is to see if one can find words for each incubator that seem to be more frequent within that incubator’s own content as opposed to a benchmark incubator (arbitrarily designated). Unfortunately, it would be hard to discern content if the analysis were done in one go on all of the 14 incubators in the data. Thus, it makes sense to split the relative word frequency comparison into two three pieces. One visualization concerns the German sample, one concerns the English sample of incubators with few articles, and one concerns the English sample of incubators with many articles. Tech House is removed from this analysis because there are only four articles from them per language, which is too low to see any patterns of association.

Impact Hub was chosen as the benchmark for the entire English sample (for both few and many article samples) in this case because it has the most articles in our sample by far (almost 300 articles). It makes sense to employ the incubator with most words written as the benchmark as the significant differences as well as similarities between all the other incubators

and the benchmark become more apparent. For the German sample, the incubator with most articles in is also chosen as the benchmark, keeping the same logic validation as for the English sample.

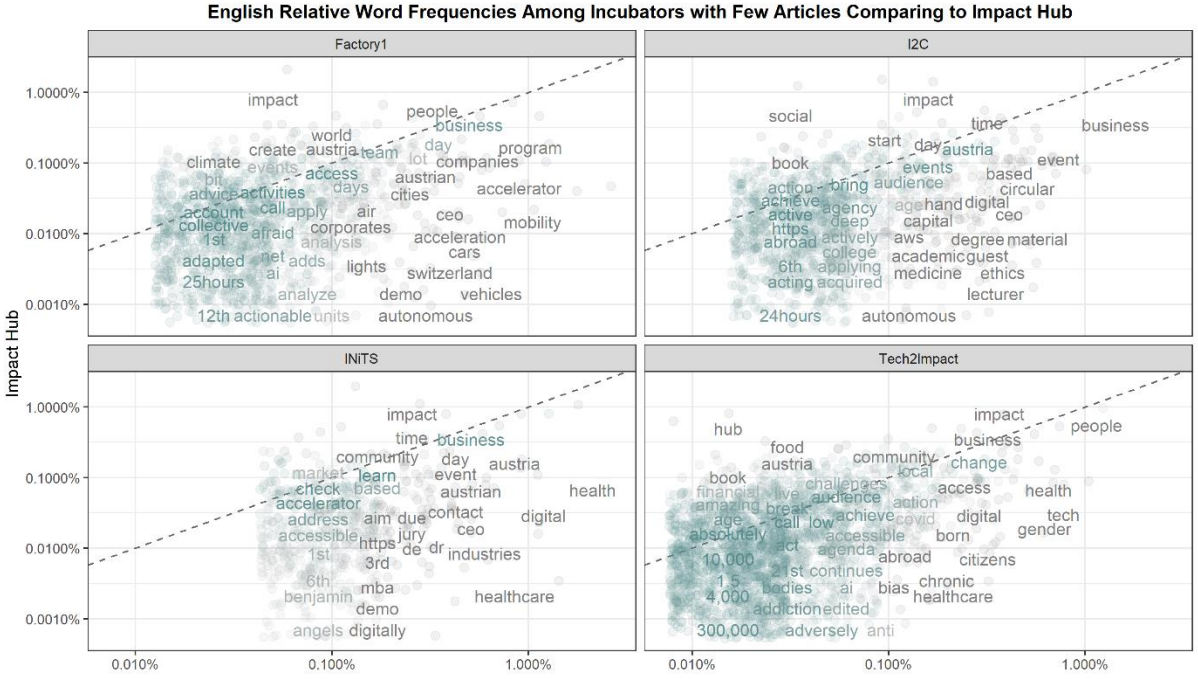

Figure 20

Looking at the comparison between incubators with few articles and the Impact Hub benchmark (most important are the dark grey words as they represent the largest difference in usage proportion between the benchmark and the respective incubator), a few interesting topics arise that are more common to the chosen incubators. INiTS, with words like ‘health’ and ‘healthcare’, distance themselves quite well from the curve, thus proving a slight “uniqueness” as opposed to the benchmark, Impact Hub. An interpretation to this would be that INiTS dedicates quite a lot of attention to healthcare start-ups, even though they have few articles written, as opposed to Impact Hub, which has a very high number of articles but not many seem to mention topics like healthcare.

Another interesting takeaway stems from i2c and their usage of words like ‘ethics’, ‘autonomous’ and ‘circularity’. This points in the direction of an incubator with philosophical and sociological themes in their articles. Lastly, the relation between Tech2Impact and democracy topics becomes apparent, especially considering their usage of words like ‘democracy’, ‘bias’, ‘government’, ‘equality’, ‘citizens’ and even ‘ethics’.

One aspect becomes rather apparent overall, however. Impact Hub does not seem to have too many (if any) words that are more common (in a way characteristic) to their content. A hypothesis here would be that with a lot of content comes more coverage of all topics and a

more general audience, thus lacking specialty terms. This can be logically and visually validated by looking at the same graph, but for the incubators with many articles.

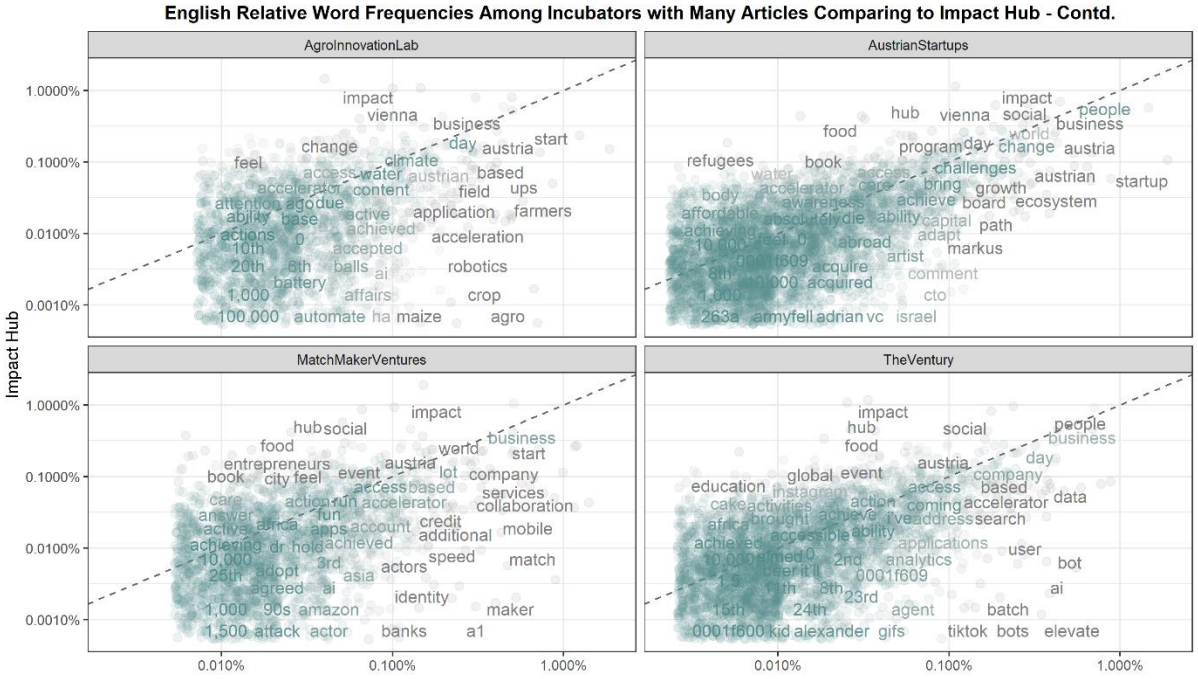

Figure 21

Looking at the comparison between incubators with many articles and the Impact Hub benchmark, the previous hypothesis seems lightly disproven here, as specialized ventures like Agro Innovation Lab seem to differentiate themselves from the benchmark with words like ‘agricultural’, ‘farmers’ and ‘crop’. It is also worthwhile to acknowledge the ‘robotics’ scattered in the field of agriculture-talk. The Ventury also differentiate themselves quite well with words like ‘elevate’ (the name of their staple accelerator program) and ‘hacking’. Looking at these distributions as well, it is safe to say that choosing Impact Hub as the benchmark was an appropriate choice as they seem to have the least unique or particular words. Consequently, the German relative word frequencies can also be explored.

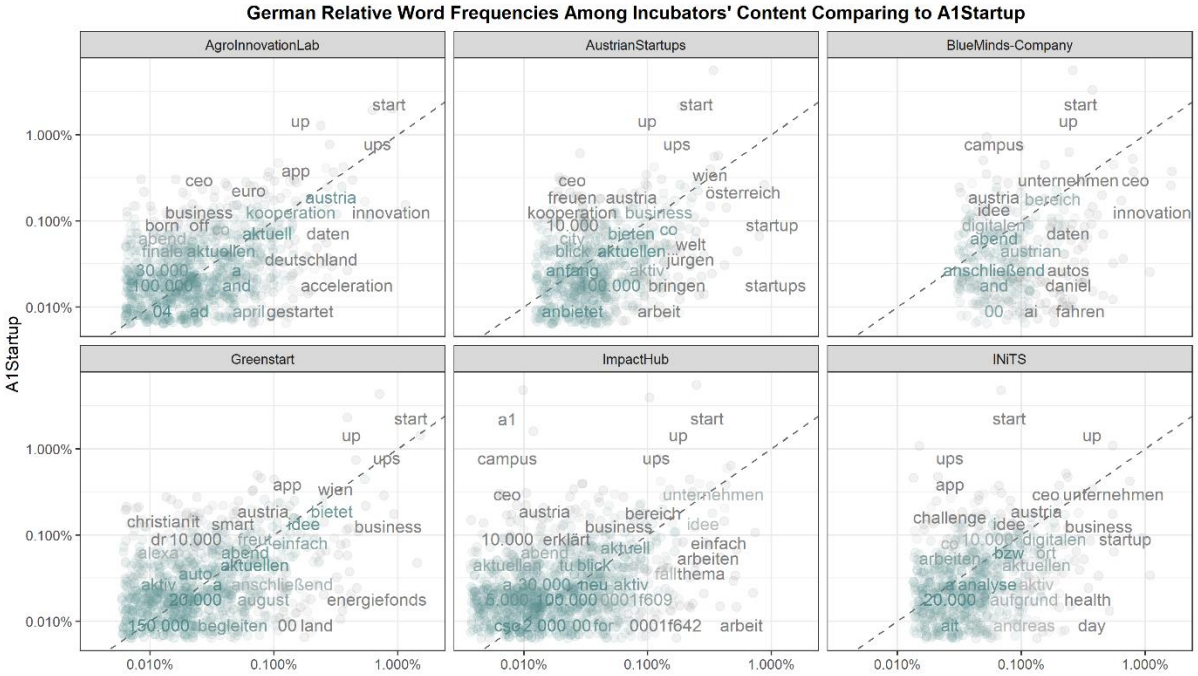

Figure 22

Unlike the English analysis, the German counterpart does not seem to be heavier (or more populated) in the bottom-right quadrant. The lack of characteristic words showcases much higher similarities between word usage for the German content of the incubators. The only partially interesting finding here is the usage of the 'energiefonds' term which seems quite natural to be leveraged by an accelerator like Greenstart.

Overall, the word frequency comparisons did not yield as interesting results for the German counterpart, but the findings do suggest that, in the broader picture, the usage of specialized terms might just distance one incubator enough from the pack. Nonetheless, the usage of this specialized language is not common even for companies like the Agro Innovation Lab. The top term comparison between Agro (a specialized incubator) and Ventury (a general-purpose incubator) found in the appendix illustrates this.

Word Correlations

On the topic of word frequency correlations, it is worthwhile to inspect whether the frequency and word usage correlation between incubators is as high as the previous might lead one to believe. One hypothesis would be that the Agro Innovation Lab will have low correlation scores with everyone else while the rest will have reasonably high correlations among the pack. These correlation checks need to be done two-fold, once for English and once for German.

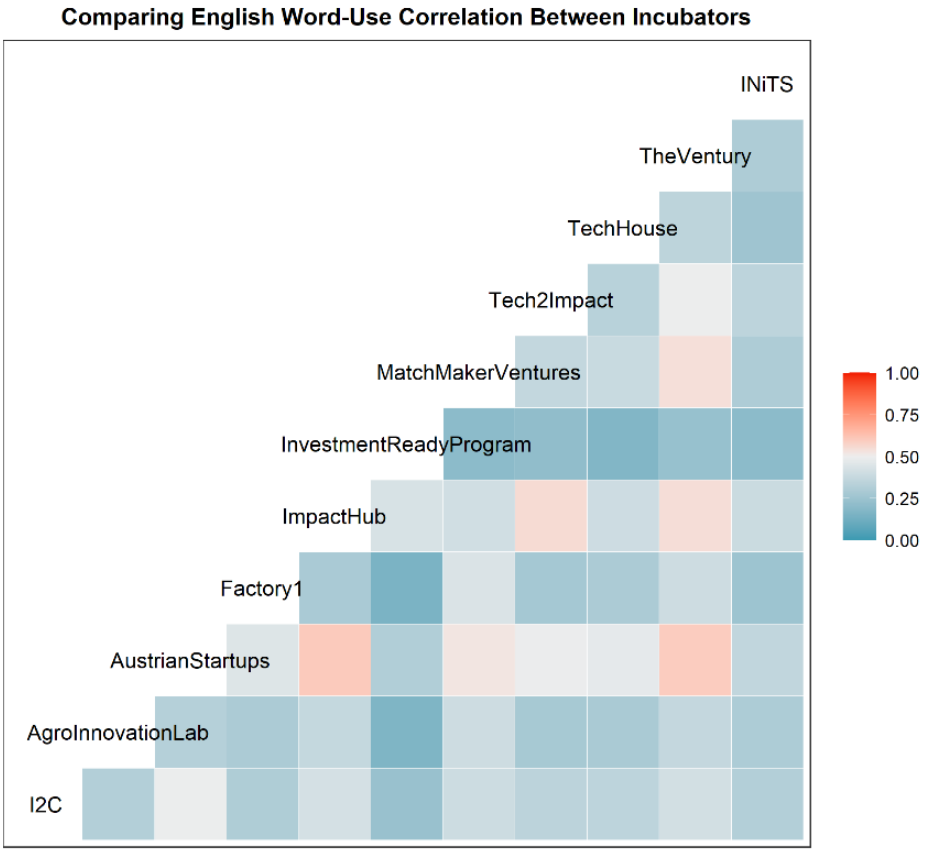

Figure 23

A few surprising findings here (but also rather depressing overall for the startup incubator ecosystem). Factory 1 really stands out with low correlation scores all around. It seems that even Agro Innovation Lab have more word frequency correlations with the crowd. On the topic of the Investment Ready Program, its inclusion into this correlation analysis fulfills a representative sample role, but the patterns of association observed are hard to interpret with such a small base of articles. With very low correlations, but also low number of articles for about half of the incubators, the German word-use correlation cannot be used for interpretation (the graph, however, can be found in the appendix).

TF-IDF

As word-use correlations seem to be too high between the incubators, it is worthwhile to look at the words that set them apart. Quantitatively, the characteristic words can be measured with the previously introduced concept of tf-idf. As the term frequency would not be high at all for Tech House and IRP's words, while the inverse document frequency would be too high, these incubators were excluded from the tf-idf token analysis. The same process applies to the tf-idf word analysis as it did for the tf-idf image entity analysis.

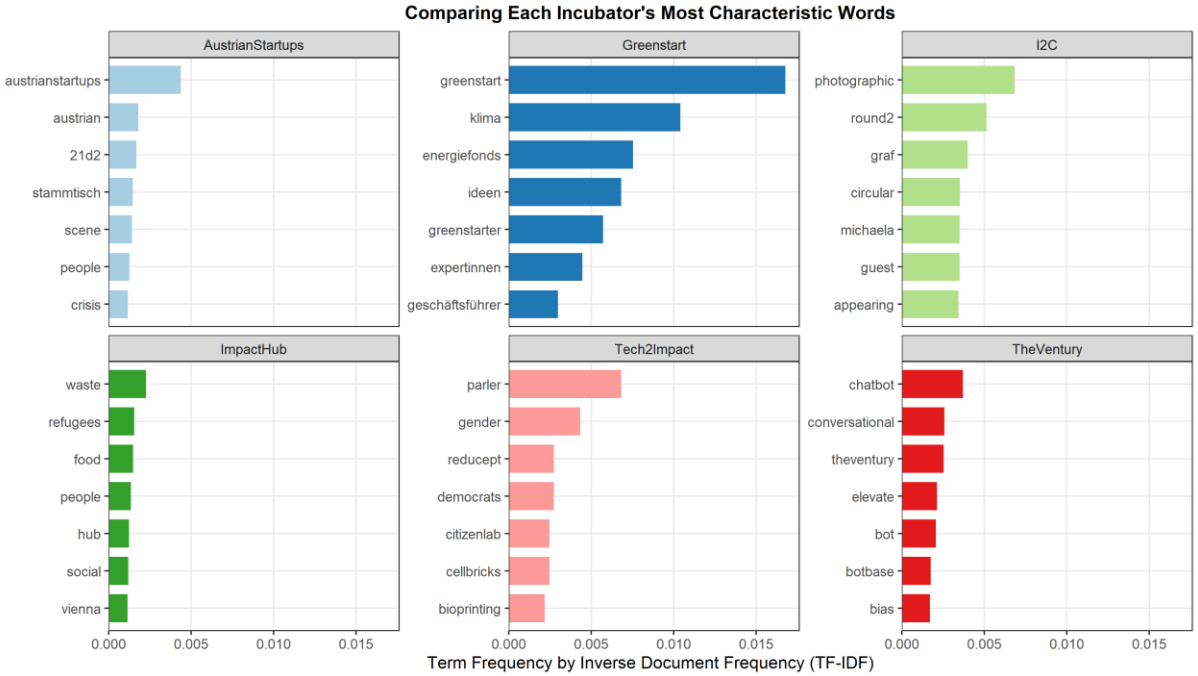

Figure 24

Figure 24 provides some useful insights on the most characteristic words per incubator. With a potential obsession for bots, biases, and communication, The Ventury stands out from the crowd. Impact Hub and Austrian Startups' low tf-idf scores across the board do not allow for much interpretation of what could ultimately distinguish their content from the pack. Lastly, Greenstart's specialization on climate and 'green' startup culture warrants them a fair number of characteristic words to help them distinguish their content from everyone else's.

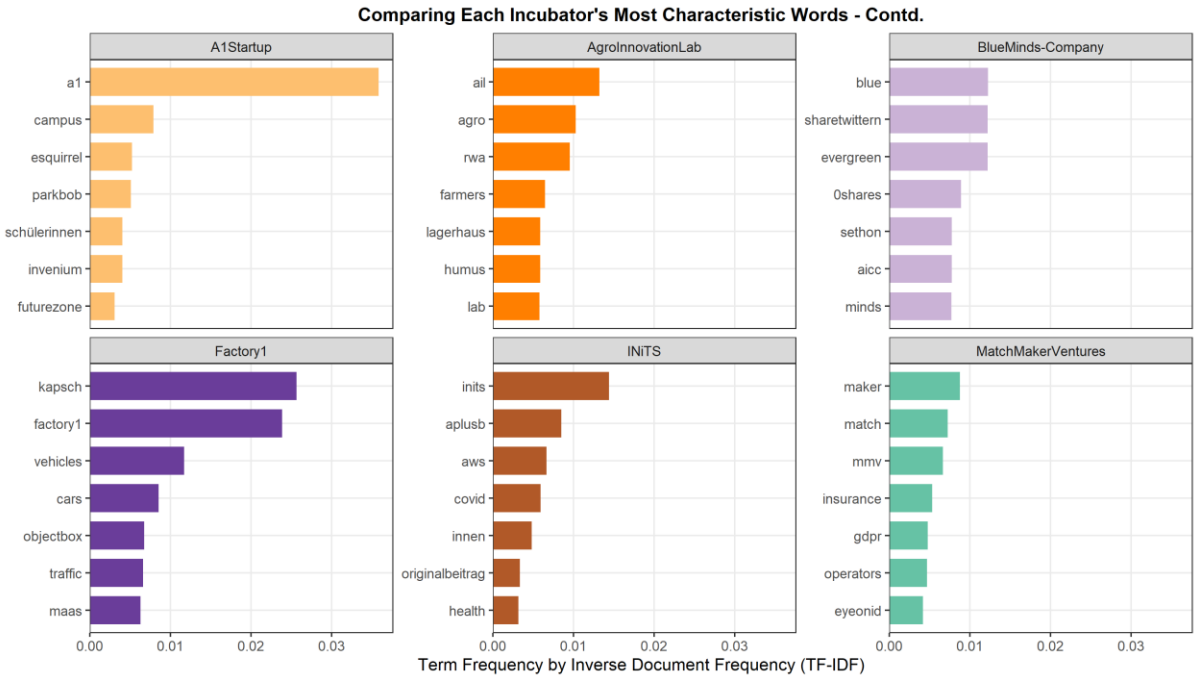

Figure 25

The findings from the relative word frequency benchmark analysis seem to prevail again in the case of Agro Innovation Lab. Factory1 also stans out from the pack with a focus on infrastructure and the automotive industry. It is important to consider that Factory1's most unique word is Kapsch. It actually refers to their CEO, Georg Kapsch, as well his large Austrian conglomerate, Kapsch Group. Curating your own content means that referencing your own parent company and name warrants you a high tf-idf, but it most likely does not translate into any content uniqueness or SEO benefits. With these tf-idf findings in mind, it is worthwhile to check if one can identify the specific incubator by the characteristic words present in its content.

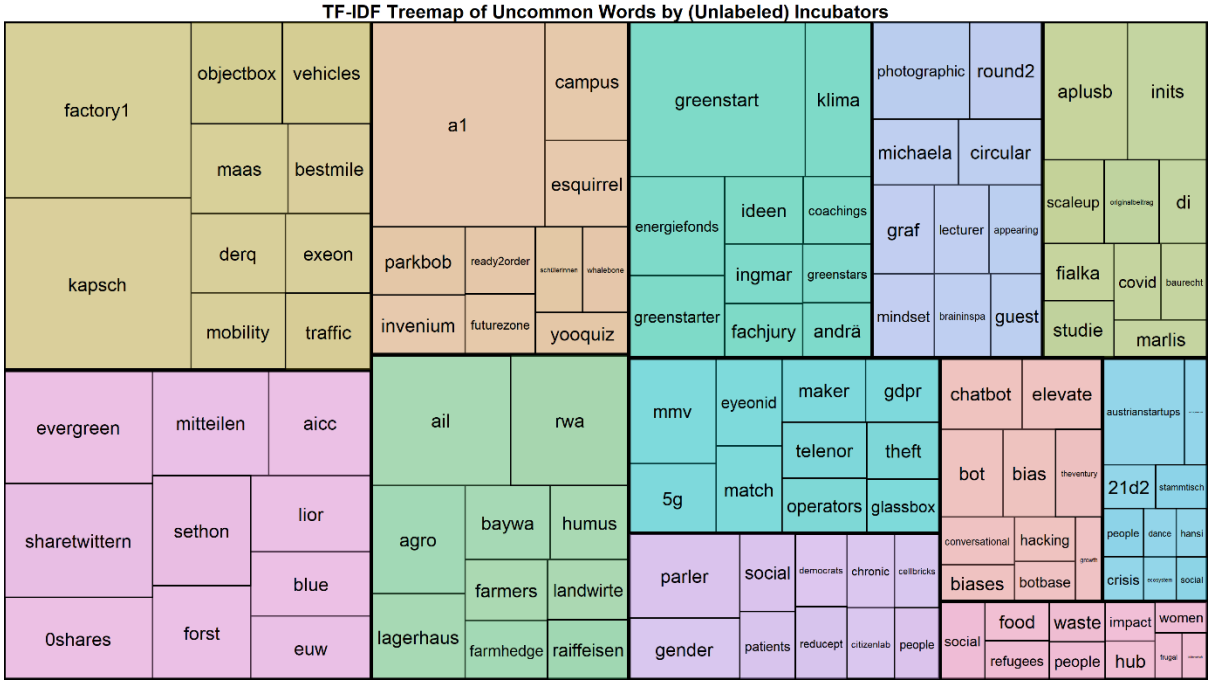

Figure 26

Figure 26 provides a slight challenge at first sight, just like the tf-idf tree map for the image analysis, but a closer look and a slightly formed eye on the previous findings and one is able to distinguish between some of the categories and attribute them to one of the 12 incubators. Ultimately, with more occurrences the tf-idf starts to matter a lot more and even more special entities might prevail.

Topic Modelling

The goal is to run topic modelling to see if articles statistically align into self-evident topics. The LDA methodology and algorithm employed here is too advanced for a thorough explanation, but the expectation is that by analyzing all the word occurrences in all of the articles, some clusters of topics can arise. This analysis needs to be done on a two-language basis as running a 2-topic LDA on a combination of articles written in English and German could result in a simple article language classification.

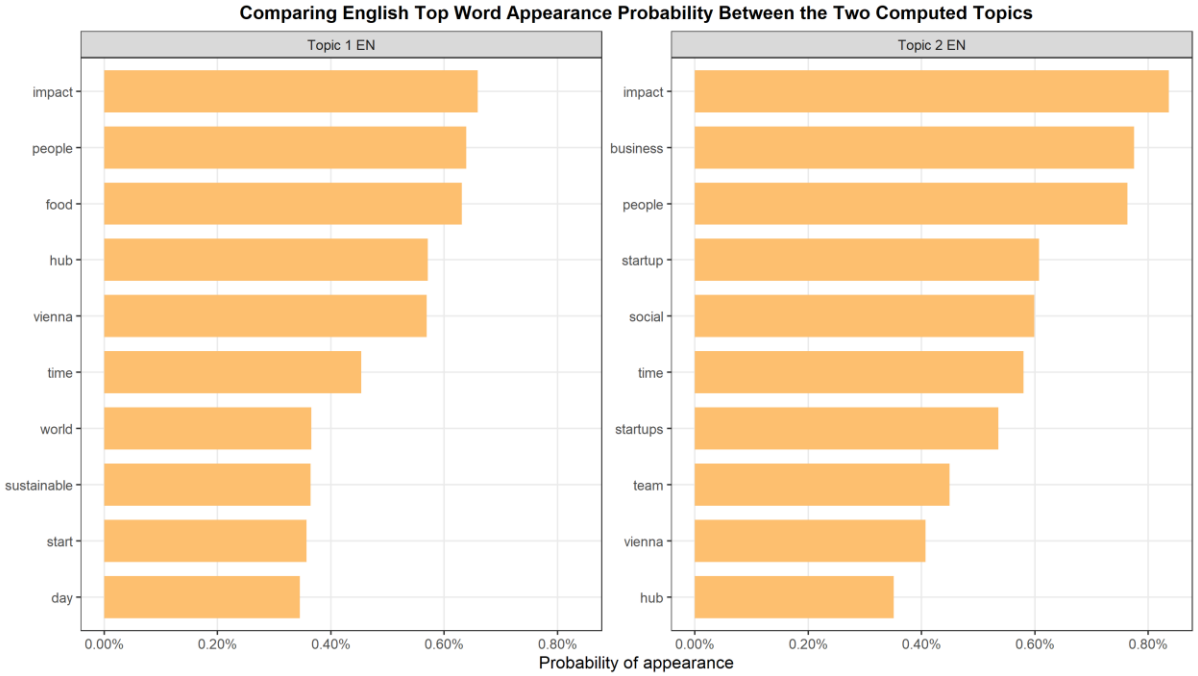

Figure 27

The same lackluster findings from the 2-topic modelling applies to the German articles. The respective visualization can be found in the appendix. What would pose more interest would be to compare the most significant disparities in likelihood of words belonging to either topic.

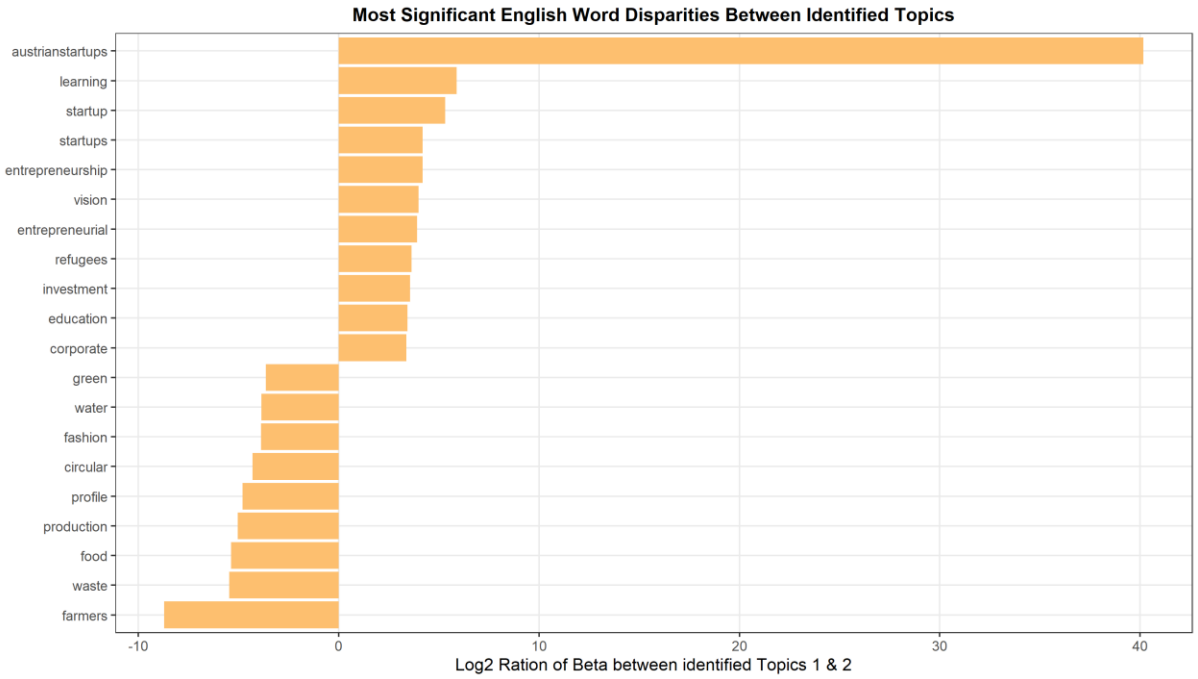

Figure 28

What this visualization aims to achieve is to show the words that are most likely to found in one topic and not the other, thus serving the purpose of uncovering the actual topics identified by the LDA algorithm. While the first topic seems to cover general elements in the

field of entrepreneurship and education, the second topic seem to be a lot more about the green circular economy. With areas such as 'food', 'waste', 'farmers', 'fashion' (one of the most polluting industries) and 'circular', it seems that topic 2 differentiates between articles about general entrepreneurship and articles about some of the world's sustainability crises. The German counterpart to this topic likelihood disparity analysis is in the appendix as the findings do not provide any insights into how the two topics might have been modelled.

LDA Incubator Classification

After comparing word disparities in topic modelling, it is worthwhile to look into topic modelling for classification. The aim is to see if some of the particular words to each incubator can help one identify the incubator producing the content. Thus, topic modelling needs to be re-run with as many topics as there are incubators, separating again by language. To ensure that the LDA classification is not affected too much by the skewed distribution of articles, only incubators with more than 10 articles were kept for this classification analysis.

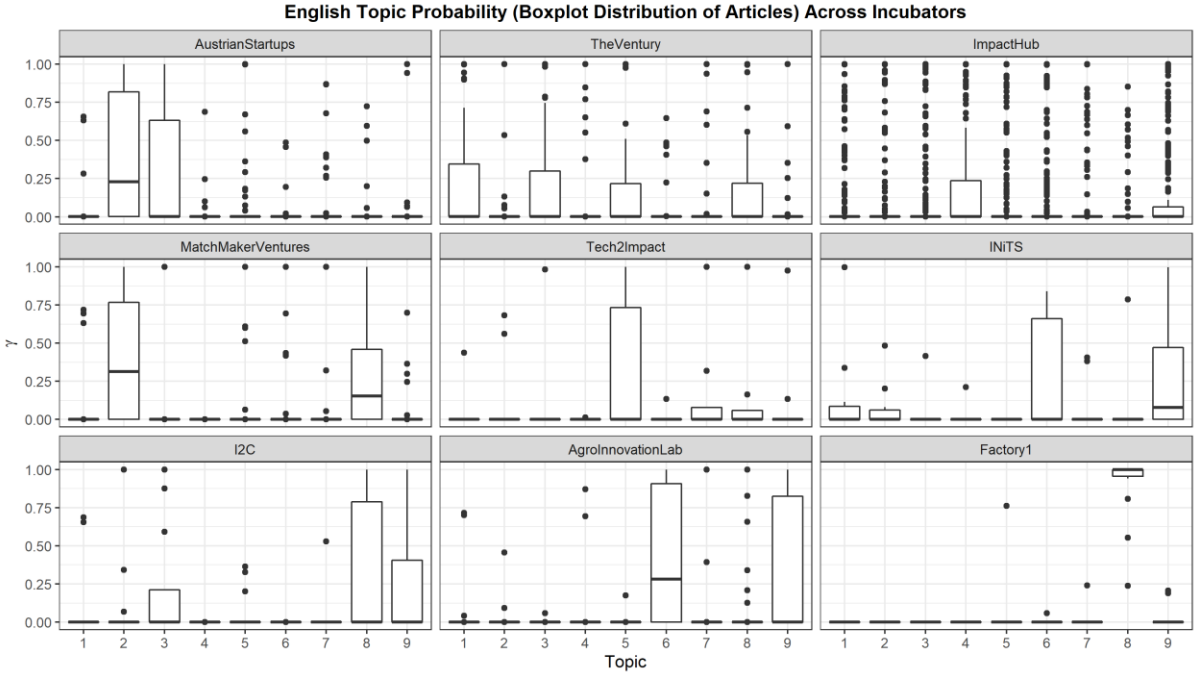

Figure 29

The findings are looking quite bleak. Factory 1, which has been identified to write a lot about its own founder and CEO, does seem to have its own main topic. What is also interesting is that each of the incubators on the bottom two rows (with around 40 articles each) seems to be mostly between two topics. INiTS and Agro Innovation Lab even seem to have the same distributions of topic probability in their articles between topic 6 and 9.

Unfortunately, it does not ultimately seem that there is much uniqueness to the incubators' news and blog contents. If LDA's black box "naturally" aligning topics do not

show more than a slight differentiation between the incubators, one could ponder at whether they all actually write about the same topics and use the same words. If the Austrian market likes this generalized content, it might come down to sheer advertisement expenditures and industry prominence between the startup incubators to lure incubatees into their programs.

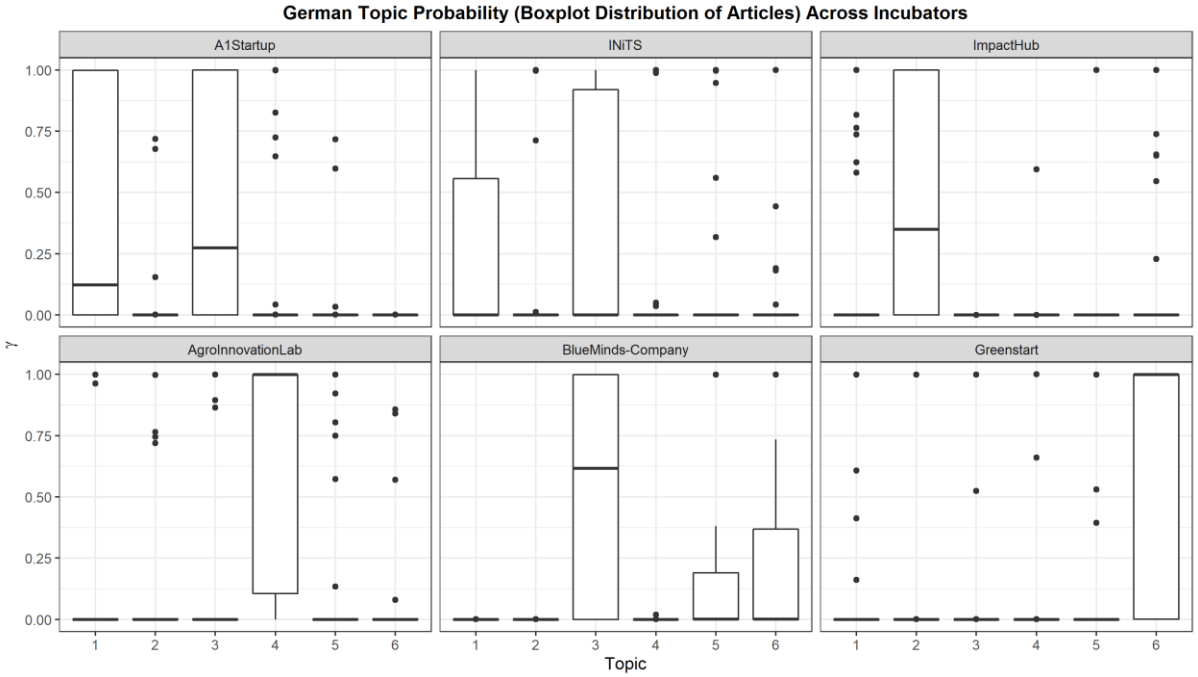

Figure 30

Similar findings are applicable to the German counterpart of incubator classification. With a mean topic 6 probability of 99% and a mean topic 4 probability of 99%, Greenstart and Agro Innovation Lab do seem to differentiate themselves from the rest. The specialized nature of these incubators and their particular topics earn them a distinguishing factor from more general-topic companies, like A1 and Blue-Minds.

Nonetheless, it is not worthwhile to look into a classification model to try to classify words by their incubators. Ultimately, this analysis non-empirically showed that incubators' content is much too similar throughout.

Word Co-Occurrences

While the single-token analysis did not identify too many differentiating factors between incubators and even topics, a multiple-token approach can be employed. The aim of word co-occurrence analysis is to gauge common word links. The rather underwhelming findings of word co-occurrence analysis on article titles can be found in the appendix.

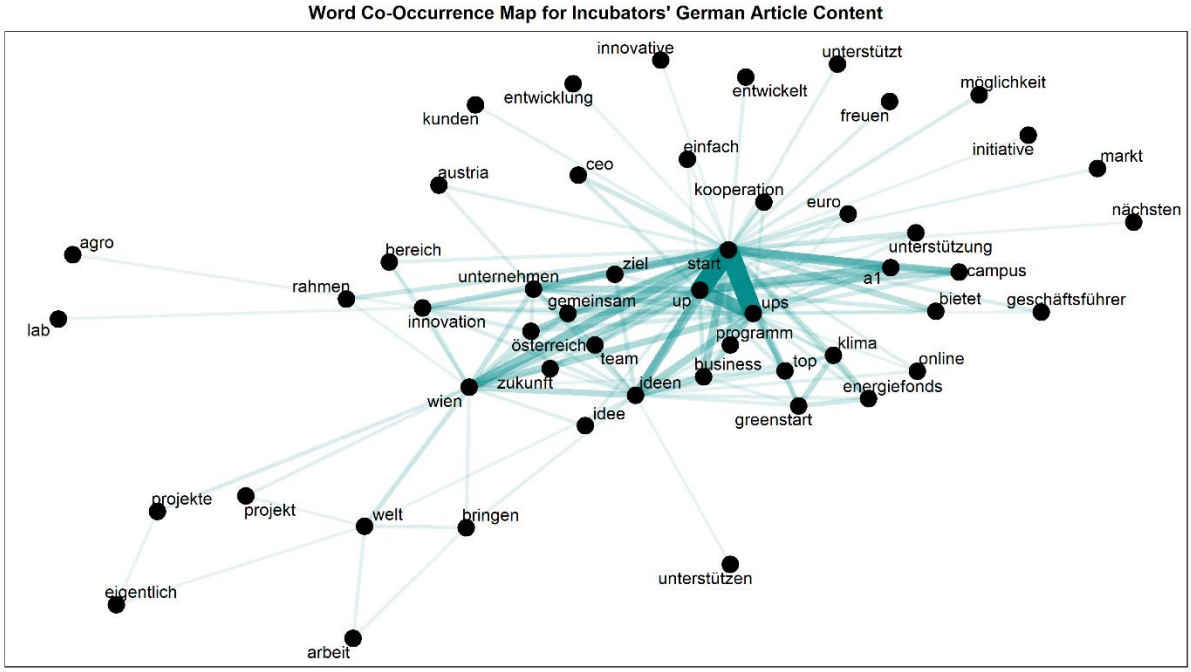

Figure 31

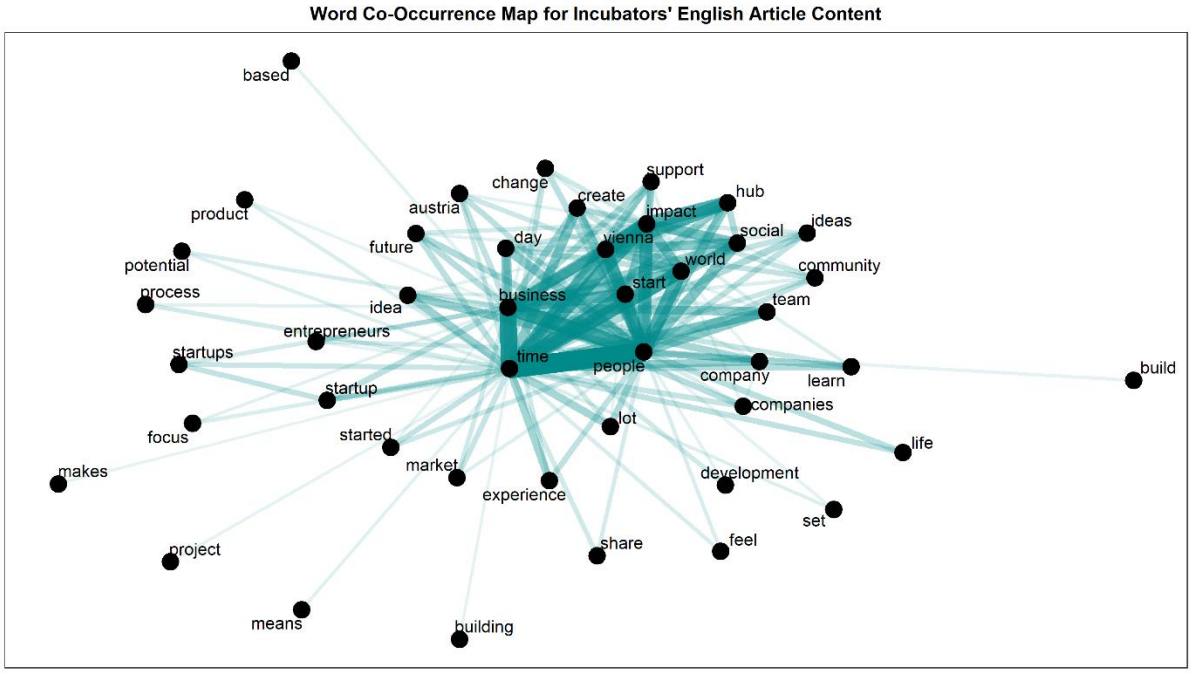

Figure 32

Unfortunately, figures 31 and 32 do not provide any real insight into the most associated words, as the majority of very popular and business-like words are present throughout. This approach of identifying common word links is plagued by the expected association of all simple words. It is better to look at word pairs and the specific pair-wise counts with n-grams.

N-Gram Analysis

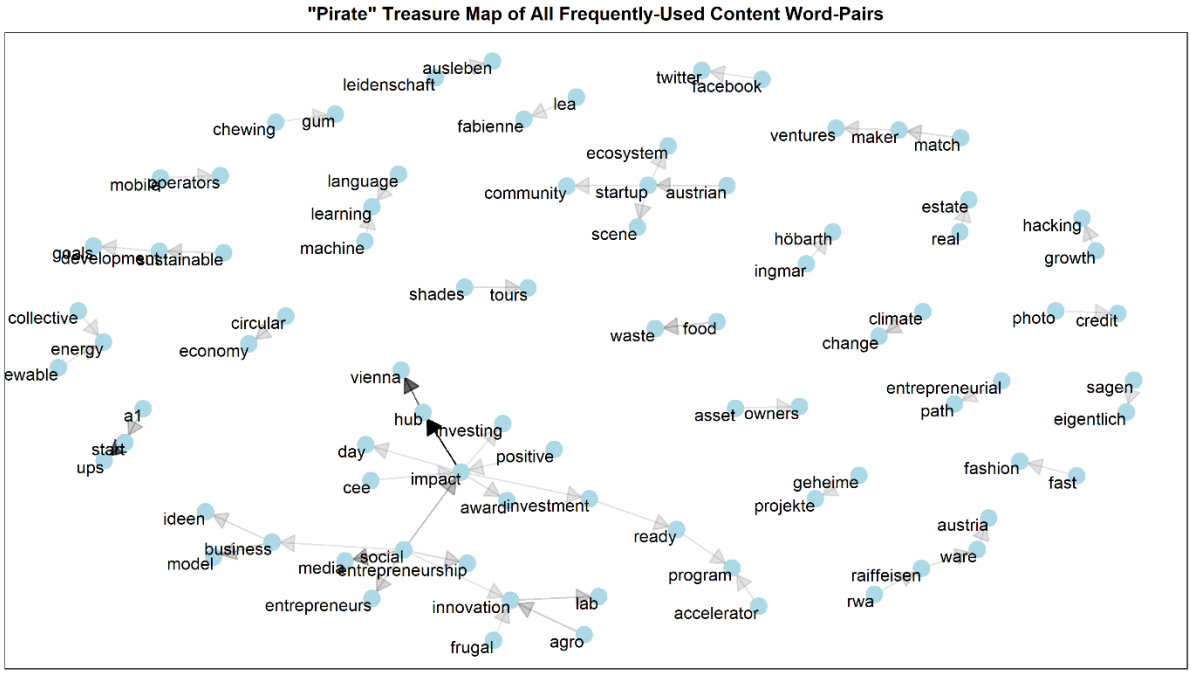

Figure 33

Figure 33 presents much more interesting associations than the word co-occurrence analysis graphs, namely because the minimum pair-wise counts were specified at 20. The common use of the ‘Austrian startup community’, ‘machine learning’, ‘CEE Impact Award’, ‘collective energy’ and ‘circular economy’ expressions gives insight into the focus of the entire startup ecosystem.

Sentiment Analysis

The last section of the EDA spans sentiment analysis, but this time at a token-level. The aim here is to identify the top words that make or break an article’s sentiment score. First off, it is worthwhile to look at the top positive and negative words identified for each language.

Word Counts by Sentiment

Figure 34: Word cloud of top positive and negative English words

Figure 35: Word cloud of top positive and negative German words

The distinction between positive and negative sentiment renders some expected results. However, because the NRC sentiment lexicon contains 10 different types of sentiment, a drill-down into more peculiar emotions will uncover the most sentiment-loaded words. The following table looks at the word counts for each NRC-designated emotion.

NRC Sentiment	Number of Words Identified
Positive	23744
Trust	12342
Anticipation	9812
Joy	7399
Negative	6188
Fear	3661
Surprise	2782
Sadness	2708
Anger	2684
Disgust	1677

Table 7

It seems that start-up incubators really love to hype people up. Out of a sample of 73,000 words, less than 10,000 really represent negative emotions and painful sentiments. Ultimately, who wants to write about the world's issues and ponder at the terribleness when one can rejoice in the novel solutions brought forth by start-ups?

The top words by sentiment for the other three lexicons: Bing, Afinn and German open source are available in the appendix. Taking a closer look at each NRC sentiment's top words:

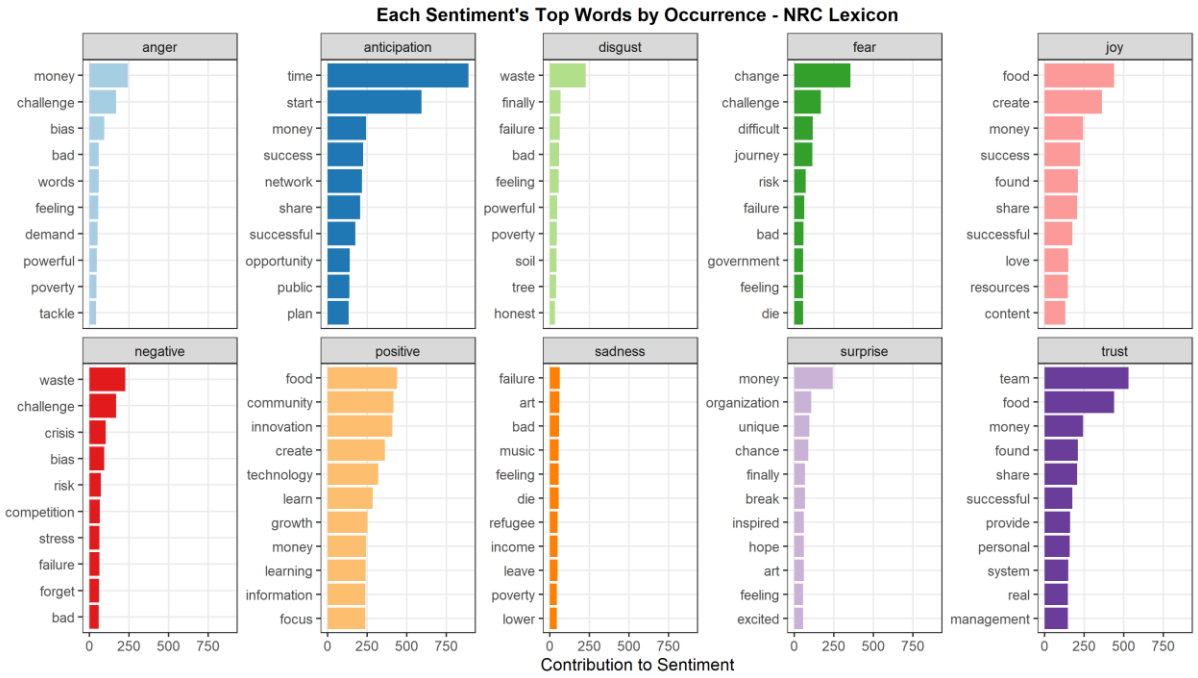

Figure 36

What is extremely interesting is that the words for anticipation are the most popular ones by far, with topics of success, money, opportunities, and funding. This ultimately shows the benefits of using a 10-sided sentiment lexicon rather than a simple positive-negative classification.

Sentiment Score Comparison – Lexicons vs. AWS

Because the sentiment lexicons employed in this analysis are all different in terms of background methodologies, measurement, scale and even sentiment denominations, it is important to see how each of them performs on the full sample of articles. This can uncover potential biases between lexicons, allowing the researcher the pick the best suited lexicon for an apples-to-apples sentiment comparison between incubators' textual content.

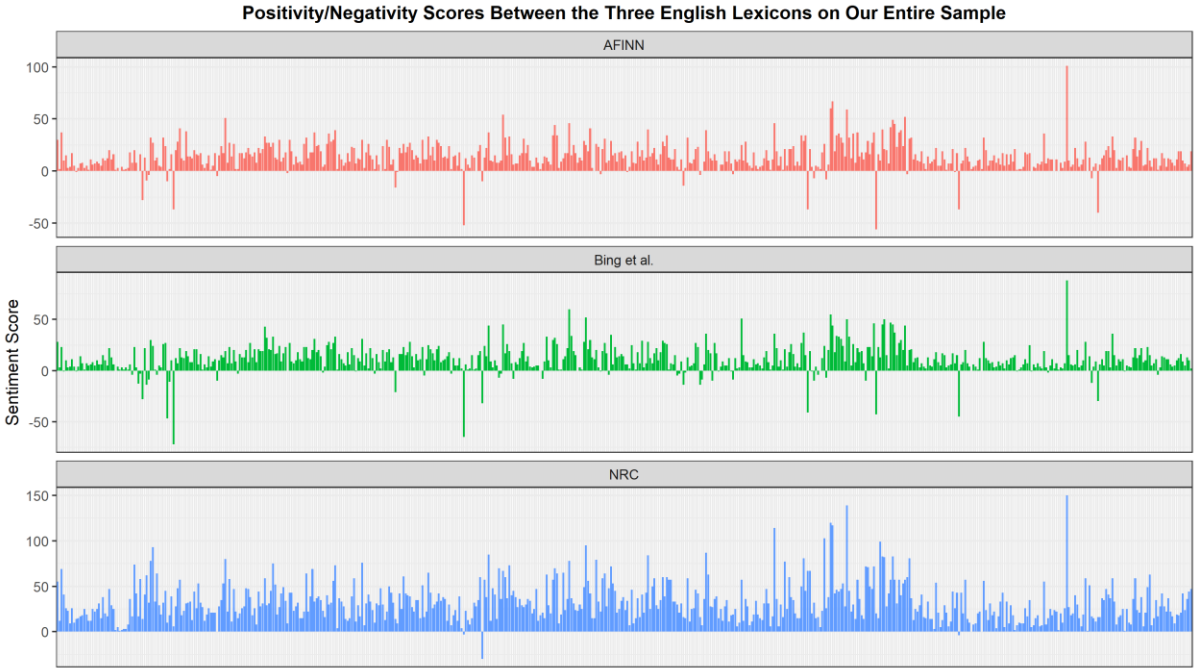

Figure 37

Figure 37 is computed on the entire English article sample, with each bar representing a single article's difference between positive and negative words. This computation allowed for the standardization of the scales, as the 10-point Afinn scale (sentiment ranges from -5 to 5) and 10-sentiment NRC scale were adapted to Bing's simplistic positive or negative denominations. Some notable takeaways from the above graph are that there might be some bias towards positivity in the NRC lexicon. Granted, all of the other sentiments were removed and only the positives and negatives were kept. Figure 37, however, allows the researcher to pick the sentiment lexicon that is most suited for further analysis. The NRC seems too biased towards positive words, while the Bing lexicon has much too simplistic denominations. The robust choice here is the Afinn lexicon. However, before continuing the analysis on Afinn, the AWS algorithm results can be included in the analysis. Without a token sentiment analysis, the AWS algorithm's black-box nature was the gold standard for measuring sentiment without contesting the results. By plotting its results against the token-analysis lexicons, the biases, and tendencies of all four (Afinn, AWS, Bing, and NRC) approaches can be compared.

Figure 38

In order to produce the graph in Figure 38, the values from the previous figure were supplemented by a neutral sentiment classification just like in AWS. The words that were not identified to be in the positive or negative categories were designated as neutral. Each bar represents the percentage difference between positivity and negativity for each article. Interestingly, AWS' positive and negative classifications are a lot more clear-cut, with a fair number of observations approaching complete positivity and a few being extremely negative. On this account, the token sentiment lexicons are a lot more muted in their positivity-negativity scores. It is curious to see if the same results apply to the German sample.

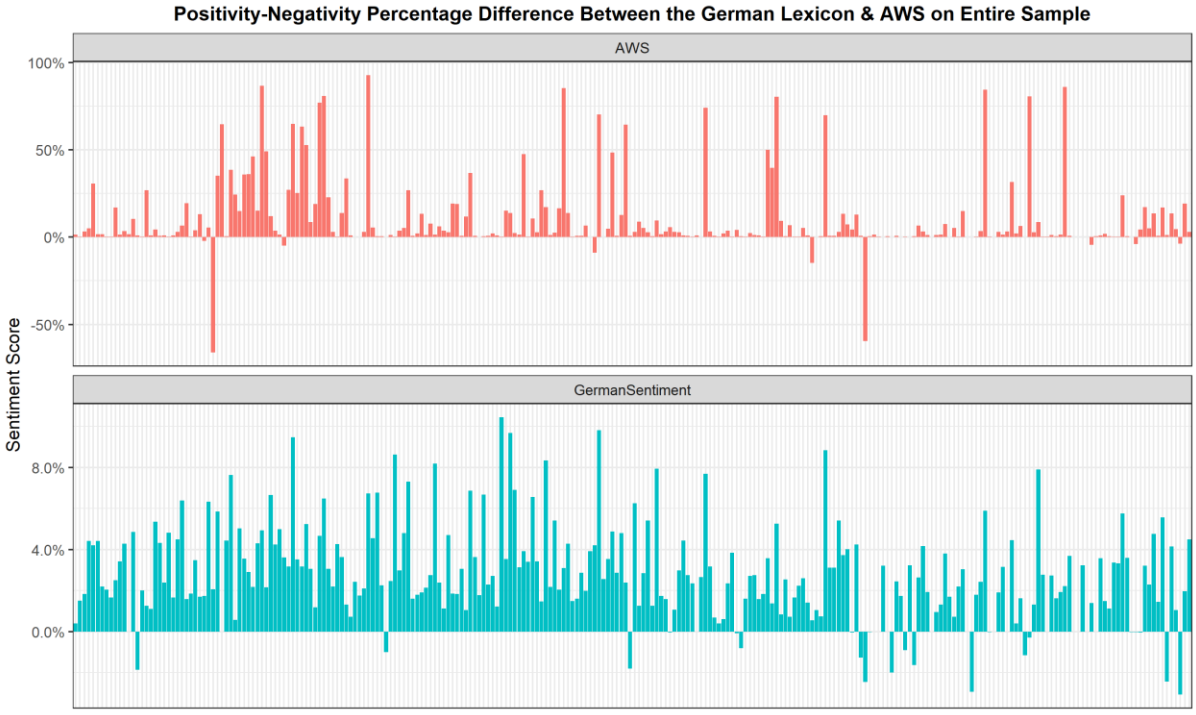

Figure 39

Overall, AWS sentiment analysis has a much more clear-cut designation of whether an article is positive, negative, or neutral. Unfortunately, the block-box machine learning nature of AWS' services means that it is impossible to know the underlying cause of its classifications.

Top Words on the Positivity-Negativity Scale

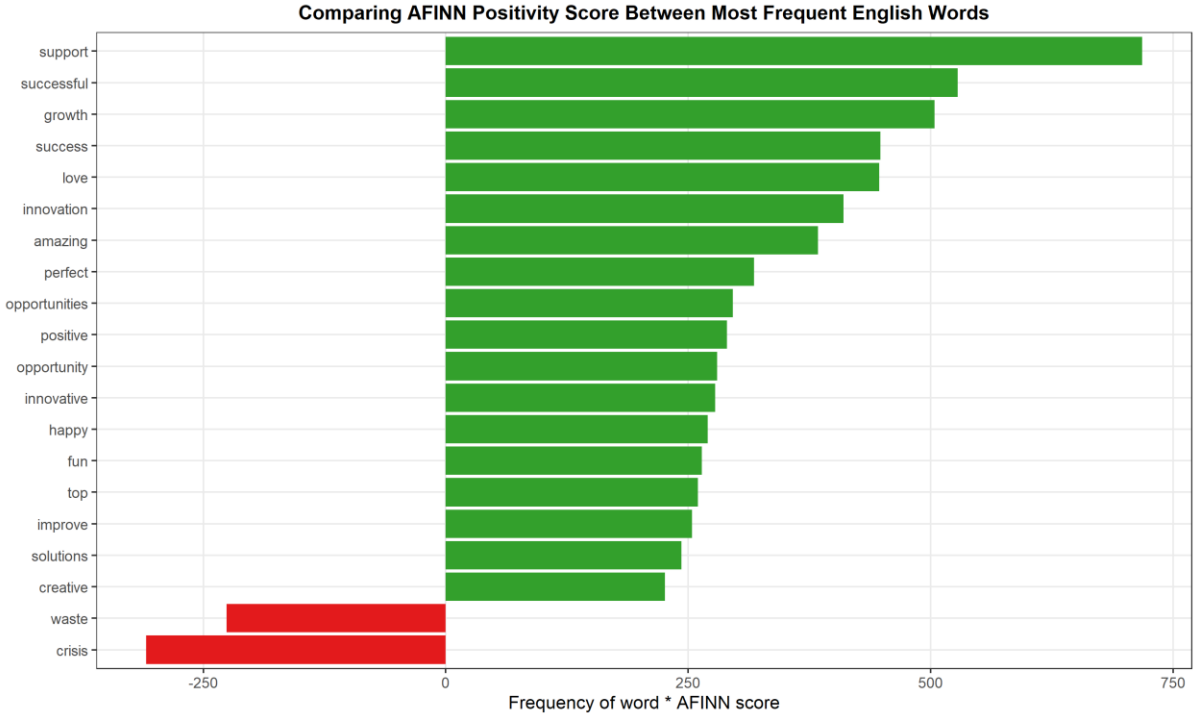

Figure 40

As stated earlier, for interpretability and robustness, the Afinn sentiment lexicon was chosen to continue the token analysis. Figure 40 relies on the Afinn score between -5 and 5 as well as a word's frequency to identify the most popular sentiment-loaded words. It should not come as a surprise that a non-conventional and non-corporate company such as a startup accelerator uses words like 'love', 'fun', 'happy' and 'creative'. In fact, the majority of popular sentiment-loaded words are overly positive. The results for German can be found in the appendix.

Introducing the Business-Oriented Sentiment Lexicon by Loughran et al.

Perhaps the token sentiment analysis has been taking the wrong approach until now? What if all these incubators actually use specialized legal, technical, and business language? To verify this, one can inspect the top words for all the business sentiment categories identified by Loughran et al.

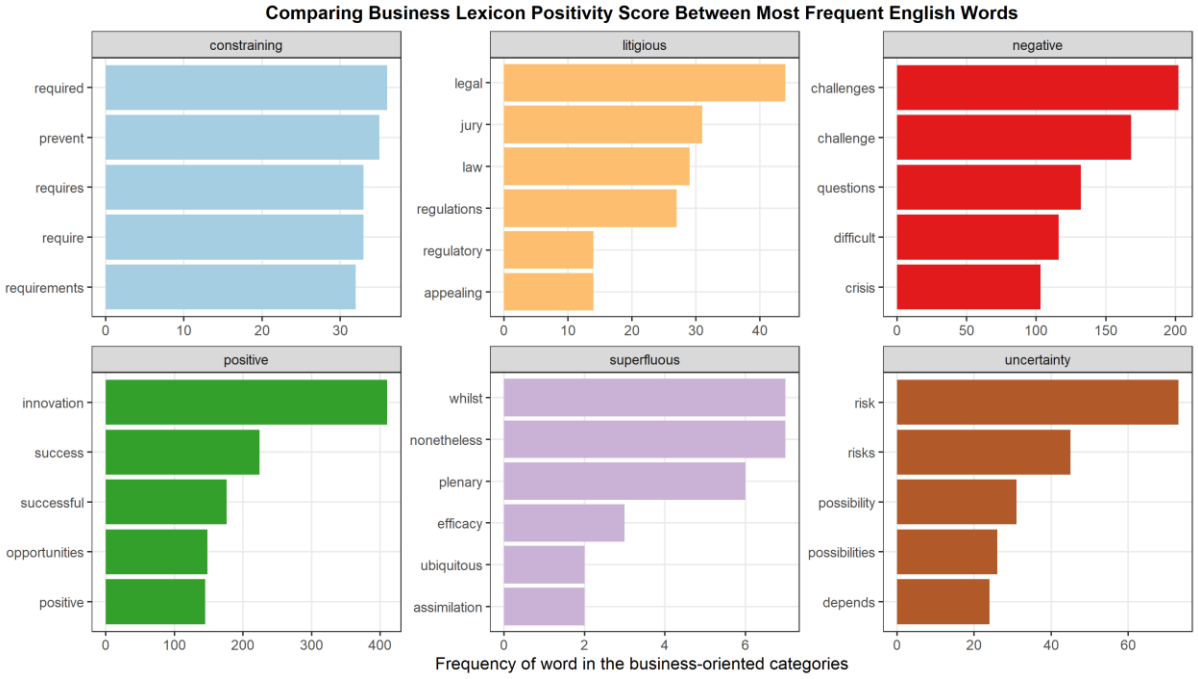

Figure 41

What this shows is that the articles written by incubators are not too specialized because language tends to just be positive or negative. There are barely any legal or business terms like 'litigation' or 'regulatory', seeing as these sentiments' top words have very low frequencies. Ultimately, it seems that sticking with a simple positive-negative denomination fits the startup incubators' ecosystem in Austria. A comparison between all incubators on their difference between positivity and negativity as part of the whole sentiment according to the Loughran et al. methodology can be found in the appendix.

Cross-Sentiment Lexicon Incubator Comparison & Results Validation

Finally, to round off the token sentiment analysis after having reviewed every lexicon and methodology, a plot of all incubators and their standardized positivity-negativity score can be made. This is done so that every reader can choose their lexicon, be aware of the biases and assumptions held by that specific methodology and contrast the score for all incubators in an apples-to-apples comparison.

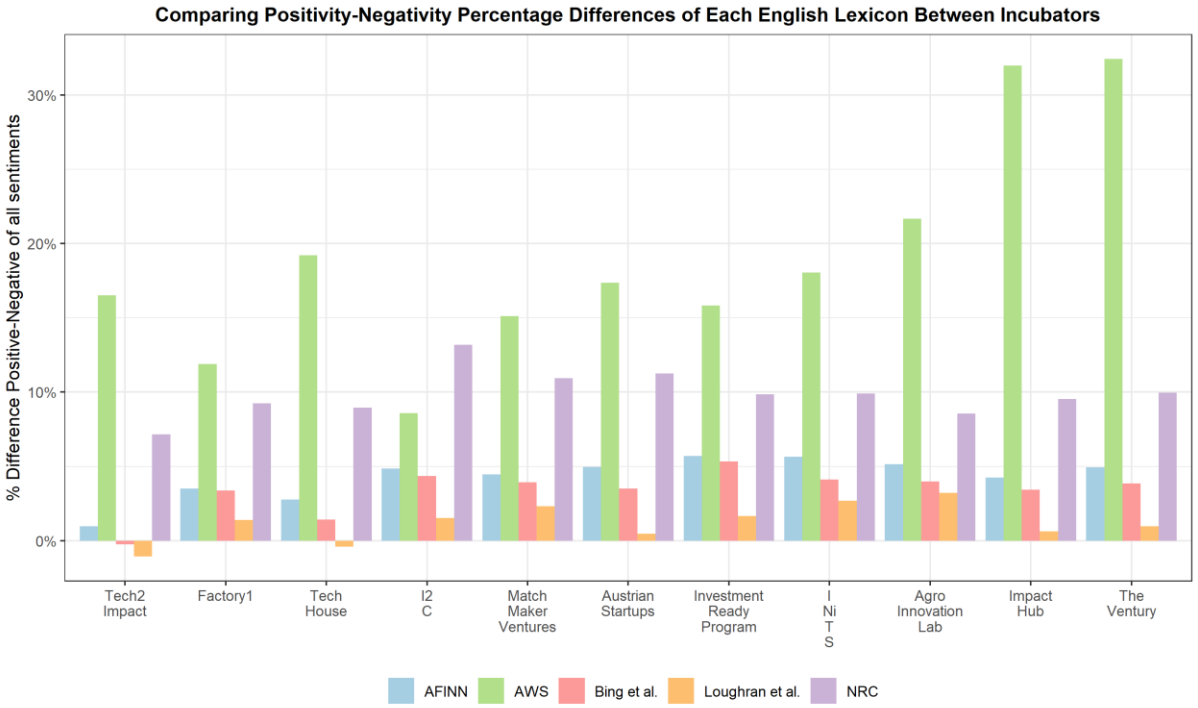

Figure 42

Figure 42 showcases that on the full sample of English articles, the Loughran et al. sentiment lexicon was by far the most nuanced, while the AWS black-box classification provided the largest sentiment disparities. Interestingly, the articles written by Tech2Impact, and Tech House had the lowest scores, even dipping into the negative territory. On the other hand, those were also identified to have a largely positive tone by the AWS algorithms.

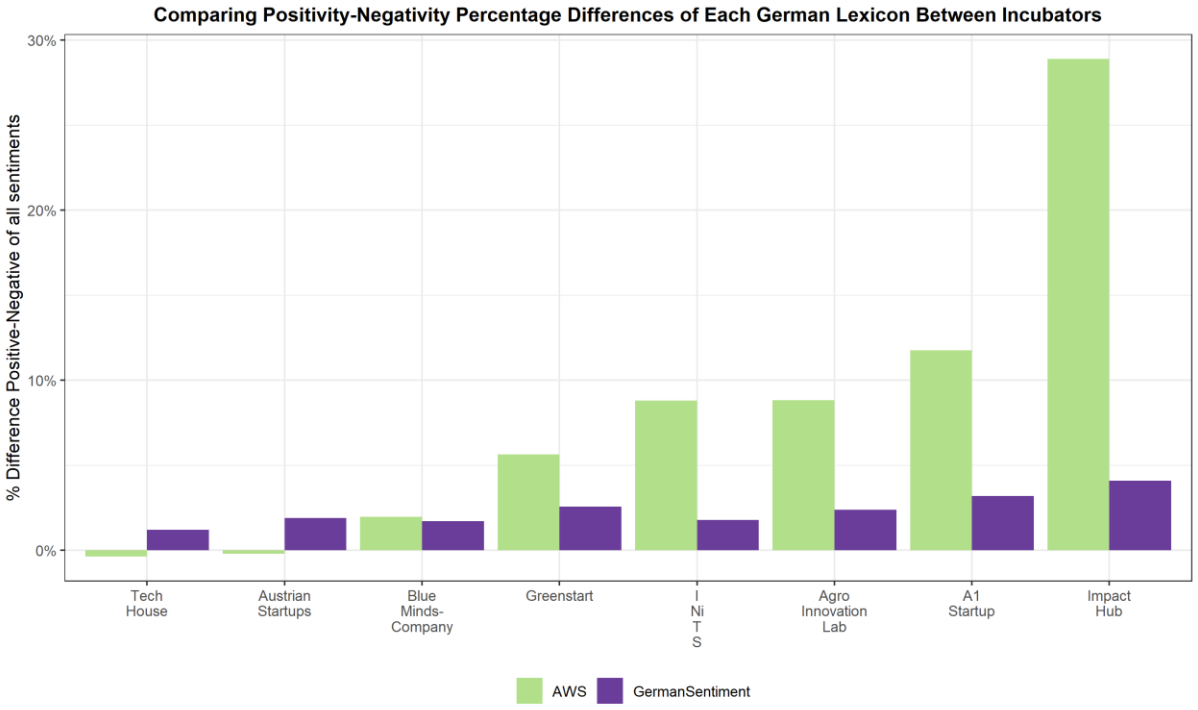

Figure 43

On the German side of the spectrum, the differences between the open-source German lexicon and the AWS algorithm are a lot more striking, with the German scores consistently hovering around 2-4% while the AWS scores range from borderline negative to very positive. The main takeaway here would be a call for finding other German sentiment lexicons to compare to the findings of this study. It is possible that the lexicon used was too simplistic for a language as poetic and complex as German.

The findings of the sentiment analysis beg for further research into the methodologies used to compute the sentiment scores and identify words. Furthermore, a deeper analysis into bigrams such as ‘not good’ and ‘not bad’ might reveal different results to this current naïve iteration of sentiment analysis.

Conclusion

The visualizations and interpretations made in this capstone, while not statistically modelled, provide a guidance framework for the client to adapt their strategy to the new market. The goal of this project was to provide an exploratory analysis of the unstructured data available about the Austrian start-up ecosystem. As the content creation efforts at the CEU iLab are largely dominated by tacit knowledge and industry experience, the client is not able to apply the same tacit skillset when embracing the new Austrian market. As such, it was of the utmost importance to try to quantify the media content patterns of the new market into explicit tidy

data that can be analyzed. The main deliverables of this capstone are not the documentations themselves, neither are the analysis' interpretations, but rather the web scraping, data cleaning, AWS augmentation and analysis scripts. These were produced in scalable fashion so further data can be gathered if more start-up accelerator companies are to be added to the initial sample described in table 5. The availability of the scraped and unstructured data also means that this analysis can be benchmarked against novel data cleaning and augmentation techniques.

Output Evaluation

The findings of this study broadly echo the tacit content creation rules that plague the unconventional start-up incubator ecosystem. When writing, curating, and sharing one's own content, it is important to consider all of the tacit decisions made, such as choice of visual content, language, writing style and brand representation. By comparing all of the articles available between the identified 14 Austrian start-up incubators, this project found that the digital content produced is extremely similar throughout, whether sentiment, key words or even image entities are concerned.

AWS Text Analysis

The AWS text analysis uncovered that the main language of choice, especially for article titles, is English. With the majority of articles containing English words, it is fair to say that popular words and phrases (buzzwords, if you will) like 'natural language processing', 'chatbots', 'automation' and 'sustainable development goals' (which were identified during the n-gram analysis) can be found in an article's title even if the body of the text is not English.

From a formatting standpoint, most articles were found to be really short and to contain a lot of sentiment-loaded words. The majority of incubators produce content that is 4-5 paragraphs in length on average.

The shortness of the content makes the AWS sentiment analysis findings even more pronounced. With a consistently positive tone, most of the identified incubators barely showcased negative emotions in their texts. A few notable examples here include Greenstart and INiTS which had low combined positive and negative sentiment scores, thus producing content with largely neutral language. This has been hypothesized to be related to the use of German. The articles written in German scored consistently lower on all emotions but neutrality in the AWS analysis. The English versus German sentiment analysis should be taken further in upcoming research to try to pinpoint what it is about the German content in this sample that gives an overly neutral sentiment (to a machine learning algorithm like AWS Comprehend; not speaking from a human's perspective).

Lastly, figure 11 showcased what could be dubbed as the most sentiment-loaded articles. These outliers had fairly large scores on the negativity and the positivity spectrums. It would be worthwhile to look into these outlying sentimental-rollercoaster articles at a later stage.

AWS Entity Analysis

The image entity analysis with AWS uncovered that the overly positive articles were linked with elements like hugs, common fruits and even artwork such as sculptures and statues. On the other hand, the predominantly negative articles were linked to common household items and to elements like locks, blades, and protests.

Most of the insight arose from the tf-idf analysis on the image entities, thus helping to identify the most characteristic image entities for each incubator. With a keen focus on infrastructure, Factory1's visual content stood out containing elements like highways, cities, asphalt, and tarmac, while Blue-Minds Company's visual content distinguished itself by using a lot of space-themed elements. These findings were hypothesized to be consistent with stereotypically motivational online content showing crowded cities, bustling with life or the universe as presenting a complex yet powerful theme. Lastly, more specialized incubators like the Agro Innovation Lab set their content apart by using entities consistent with their agricultural theme, such as 'soil', 'countryside', 'green', 'animal' and 'vegetation'.

Token Text Analysis with R

For this last section, the results point towards an undeniable similarity across the board of Austrian start-up incubators' textual content. The majority of buzzwords were found to be employed almost equally as much by each incubator. Especially concerning the English language articles, a 50% correlation across the board of incubators seems to be the norm. This is what ultimately led the research into trying to identify each incubator's characteristic words.

The tf-idf analysis did not uncover very specific words and themes in the incubators' content, but rather showcased that the only company setting itself apart was Factory1 because of their constant references to the CEO, founder and business conglomerate owner, Georg Kapsch. The search for words that set the start-up incubators apart continued by employing the LDA algorithm to try to uncover patterns of associations between all the articles. The findings here did pose an interesting point, which is that, with a 2-topic model, the more general entrepreneurship content is put into one group while the green circular economy and sustainability themes were placed in another group. This interesting finding begs for a further

exploration of topics with LDA. The incubator classification with topic modelling attempt ultimately showcased that most of the incubators' content contains close to the same words. On the word co-occurrences and pairs analysis level, most of the findings were rather simple, with the exception of a few key phrases. Figure 33 identified the common use of the 'Austrian startup community', 'machine learning', 'CEE Impact Award', 'collective energy' and 'circular economy' expressions, thus giving insight into the focus of the entire Austrian startup ecosystem. The key takeaway from the single and multiple-token analyses is that searching for explicit patterns of text in an ecosystem otherwise "plagued" by tacit knowledge, industry acumen and networking knowledge will ultimately single out the specialized players (such as Greenstart and Agro Innovation Lab) and largely ignore the contributions of the incubators producing a lot of general-purpose content (such as A1 and Impact Hub).

Finally, the sentiment analysis helped identify the major sentiment-loaded words, especially for emotions like anticipation and trust. The comparison between all the sentiment lexicons used meant that the reader (or any follow-up researcher) can choose to follow the assumptions and methodology of any of the six (between English and German) lexicon analyses. Ultimately, the sentiment analysis showcased the key differences between running single-token-based emotion detection as opposed to AWS' advanced machine learning algorithm. The findings from the token-based approach do reinforce the AWS ranking of start-up incubators by the positivity of their articles. The Ventury and Impact Hub were found to have the most positive content throughout.

AWS Costs – Estimation

While the data gathering and analysis scripts were run on the researcher's local machine, the data augmentation scripts made repeated API calls to the Amazon Web Services machine learning functions. Those services charge a very small fee per query. A query can be understood here as a single text or a single photo. The following table breaks down the incurred costs of using each AWS machine learning service for this capstone.

Service	Request	Query	Cost (USD)
Comprehend	Language Detection	1100 x 400 characters	$0.44
Comprehend	Sentiment Detection	800 x 4000 characters	$3.20
Rekognition	Image Label Detection	3700 images	$3.70

Table 8: Amazon Web Services – Estimated Costs

Limitations

As for the limitations of this study, they span two-fold: methodologically and technically. On a methodological level, the assumptions made about the start-up incubator ecosystem comprise the sample representation and the narrow focus on self-curated and self-distributed content. The incubators' self-distributed content also spans social media websites such as Facebook, Instagram, or Twitter. Thus, while the sample in this study might be representative of self-curated blog content, it is definitely not representative of all content. By excluding social media content from this analysis, some of the identified start-up accelerators were not included as they do not have proprietary blog content. An expansion of this study to social media would validate the methodology and lessen the assumptions.

On a technical level, the R scraping scripts might be scalable and they could scrape even with failures, but when the websites change structure, they lose all functionalities. The limitation here stems from the way the *rvest* package is written, namely that HTML nodes need to be leveraged so that content can be scraped. A better alternative (to search for) would be a package that does not need to drill-down into the specific HTML nodes.

Another technical limitation is that the sentiment analysis with tokens was merely done on single tokens, without taking into consideration more complex sentence structures such as negations, double negations, or composite words (of which there are a lot in German).

Recommendations

The resounding recommendation of this study is to extract all of the tokens, phrases, entities, and emotions of interest to leverage in the client's content. These elements and their performance can be evaluated externally by doing A-B testing using SEO tools. As such, the iLab is encouraged to create content to test on the Austrian market following A-B testing rules, publishing certain articles and images using this analysis' identified common phrases while also publishing content relying entirely on tacit knowledge. The success of both types of content as well as keyword prominence can be measured later with SEO tools such as Ubersuggest. It has been agreed with the client that this work will be done at a later stage by the iLab's SEO and marketing specialists.

The rest of the recommendations are all calls for further research and a continuation of this project with more data. As discussed previously, if used with more scale, the AWS machine learning services' outputs as well as tokens extracted can identify even more characteristic words and entities through tf-idf analysis. By supplementing the current dataset with more content from the incubators' websites, such as descriptions, team introductions and other

promotional visual content than what is shared in the blog posts, the analysis can be expanded to cover the entire website content spectrum. Of course, the next addition would be the social media content (accessible through paid APIs) that is shared on Facebook, Twitter, Instagram, and LinkedIn. These two data supplements would make the sample a lot more representative of the entire population, that being the entire Austrian start-up accelerator ecosystem.

On a technical level, a call for further research into AWS Comprehend and Rekognition's features, such as text entity analysis and famous people image recognition is strongly made. This project showcased the untapped and cheap potential that AWS machines learning services provide instead of running time-consuming and computationally expensive data analysis techniques on one's local machine, like the case of this project's R analyses.
For self-evaluation purposes, the scraping and pre-processing scripts can be used on the iLab's own content and merged into the base data frame. It would be highly recommended that the iLab scrape their own content and run the analysis with their tokens, entities, scores, and measurements as the benchmarks to compare the Austrian incubators to.

Finally, as the data extracted for this project is open-source and the findings are not proprietary, a potential recommendation would be that the iLab promote further analyses on the base datasets provided in this project's GitHub repositories. By leveraging a public call to action through online challenges and further Capstone project proposals, the iLab can consistently improve on this project's data as well as the insights stemming from it.

Personal Note

Finally, I would like to thank the iLab team for allowing me to run a full-scale independent analysis on self-gathered data. The autonomy that they provided me with was extremely useful, as I was able to apply all the data gathering and pre-processing techniques that I have learned throughout the past academic year. Ultimately, this capstone built upon the data munging and scraping (the Coding classes) covered in the first term, AWS services (the Data Engineering classes) covered in the second term and the text analytics (the Data Science 3 class) covered in the last term.

I was able to capitalize on my web scraping fondness and recently discovered interest for unstructured text analysis in a real-world environment. The project provided great insights into the tidy text approach and motivated me to read Hadley Wickham's book on the R tidy universe. I was also able to get independent exposure to the AWS machine learning suite.

On a portfolio level, I have learned to produce data visualizations for a crowd of non-tech-savvy people and to deliver charts with clear, concise, and simple, but effective measurements. Furthermore, creating reproducible web scraping scripts taught me how to templatize and maintain scalable functions. This was especially useful when I had to supplement the scraping scripts with image downloading functionality. Lastly, I created and maintained a GitHub portfolio of efficient data gathering, munging, augmentation, and analysis scripts to add to my overall digital data scientist footprint.

References

Open-source sentiment lexicons:

- Chen, Y., & Skiena, S. (2014). Building Sentiment Lexicons for All Major Languages. *ACL*, (2), (pp. 383-389).

R *tidytext* packge providing English and German stop words:

- Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, *1*(3), 37.

R *afinn* sentiment lexicon:

- Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*, pp. 93-98.

R *nrc* sentiment lexicon:

- Mohammad, S., & Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, *29*(3), pp. 436-465.

R *bing* sentiment lexicon:

- Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM

Appendix

AWS Sentiment Analysis

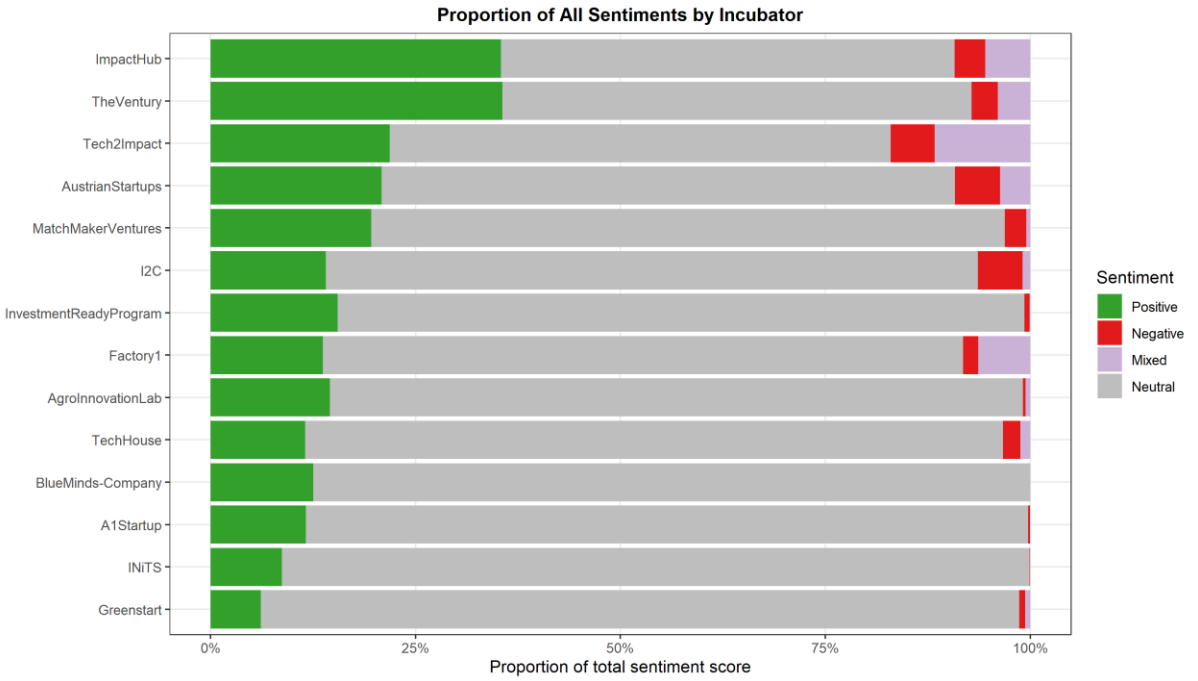

Figure 44

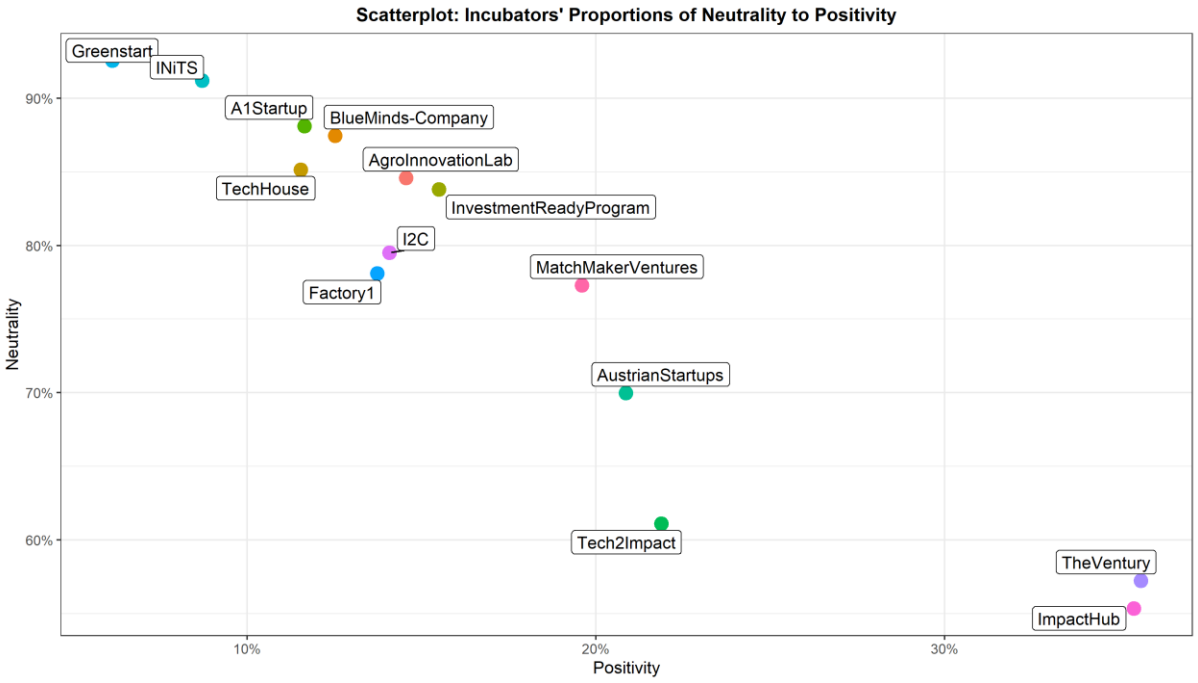

Figure 45

Word-Use Correlations

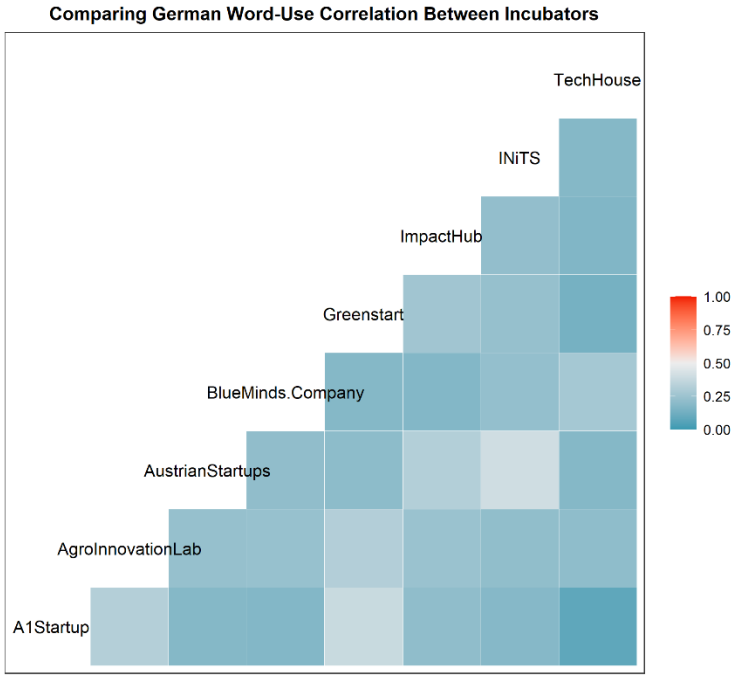

Figure 46

2-Topic Modelling for Articles in German

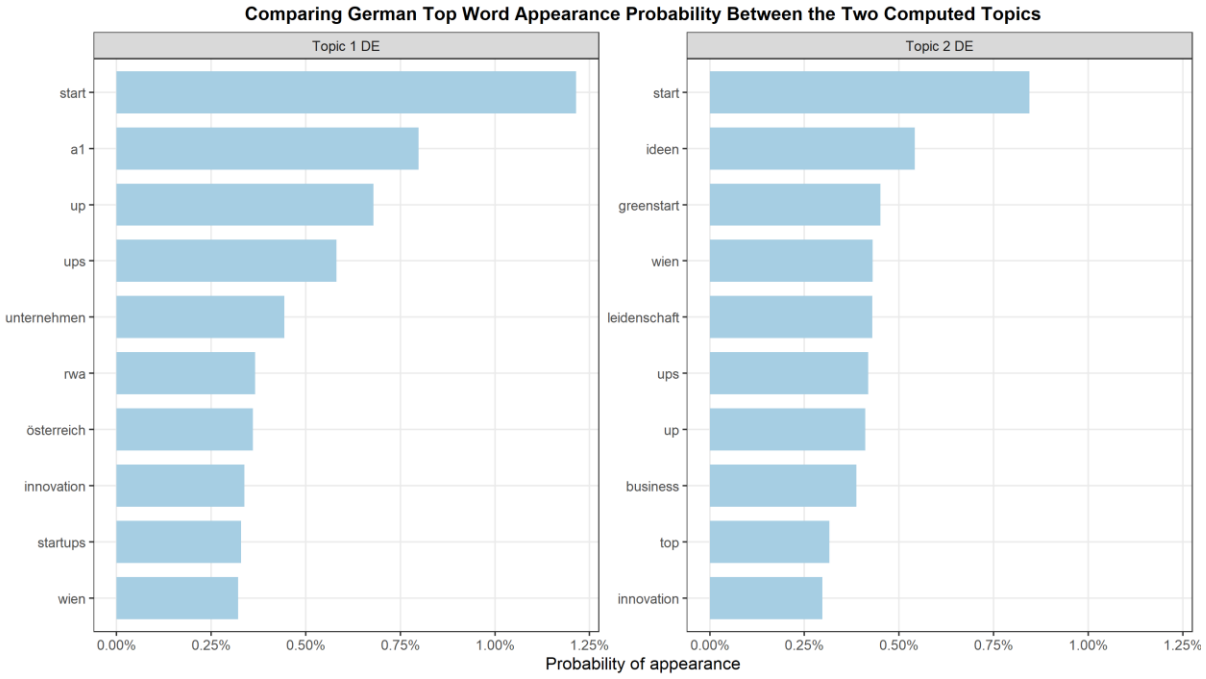

Figure 47

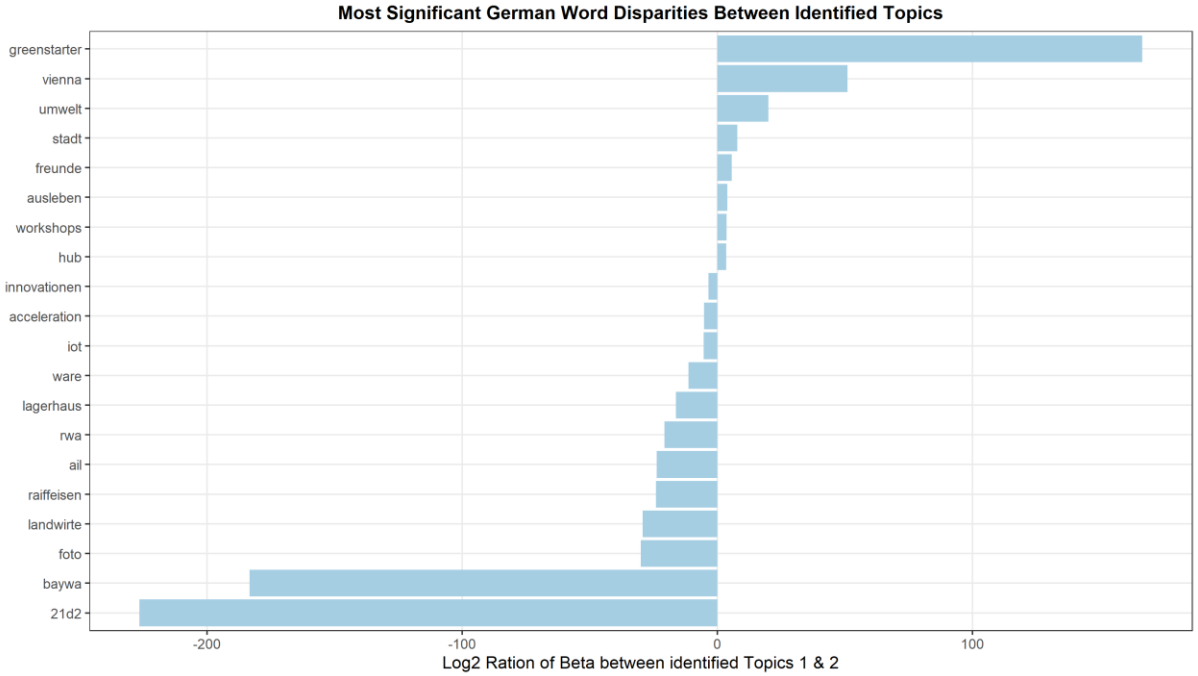

Figure 48

Compare Frequent Words - Specialized vs. Non-Specialized Incubator

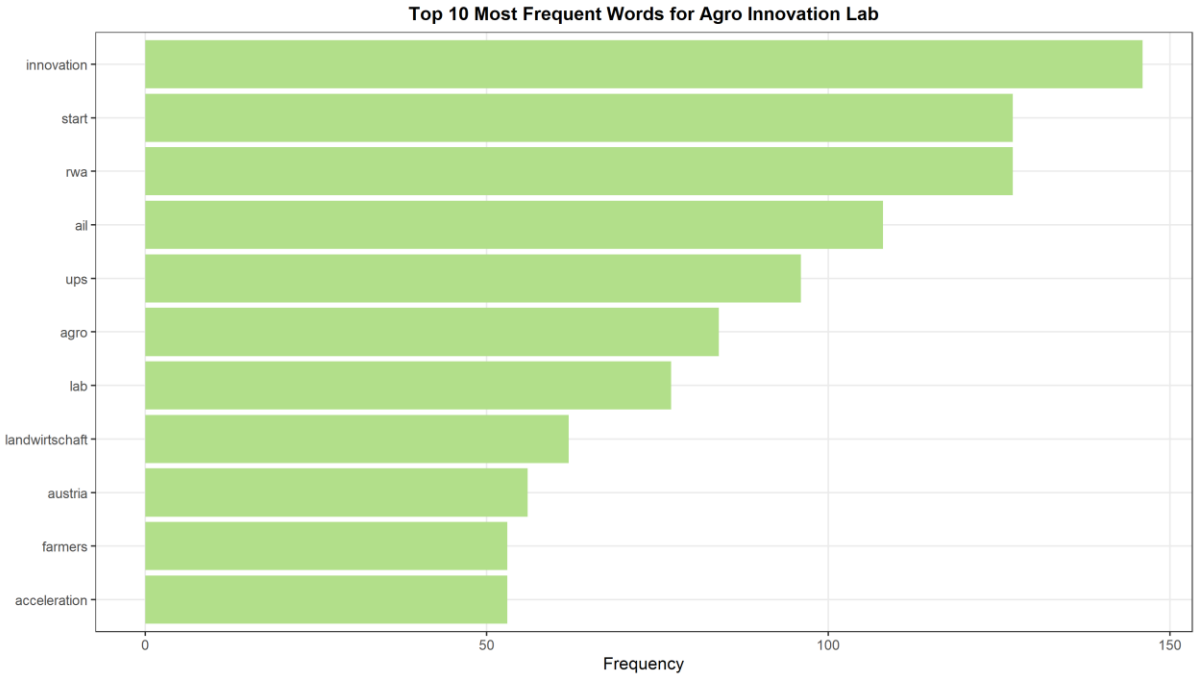

Figure 49

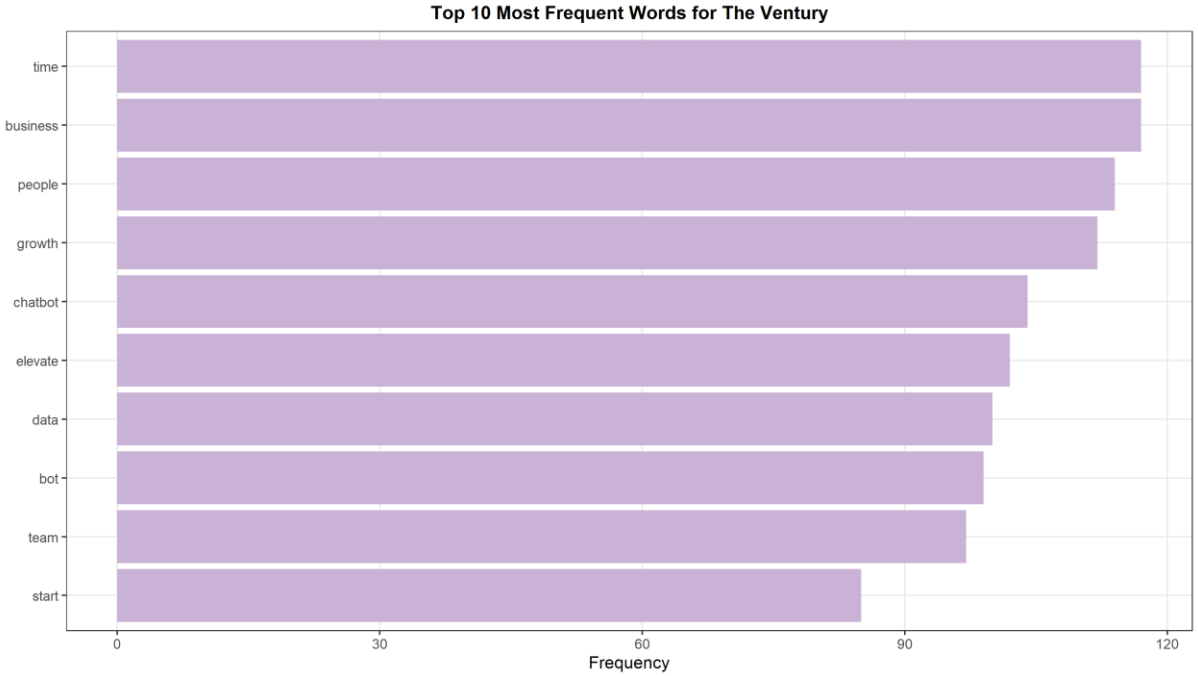

Figure 50

Word Co-Occurrence Analysis – Titles

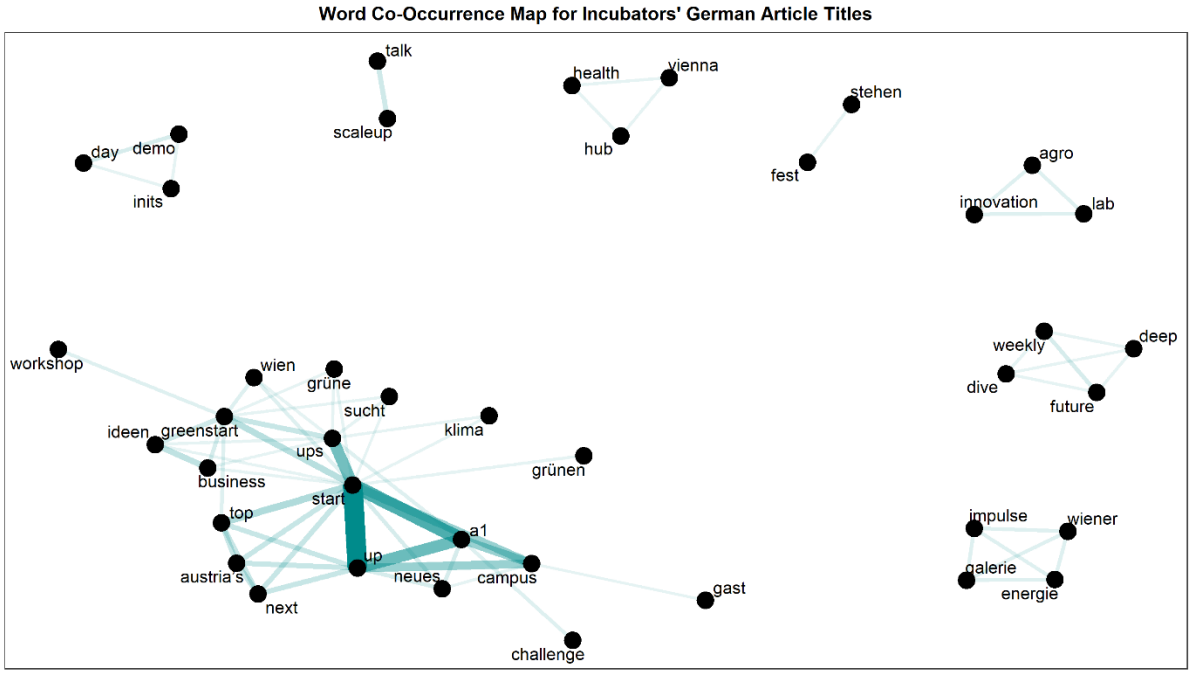

Figure 51

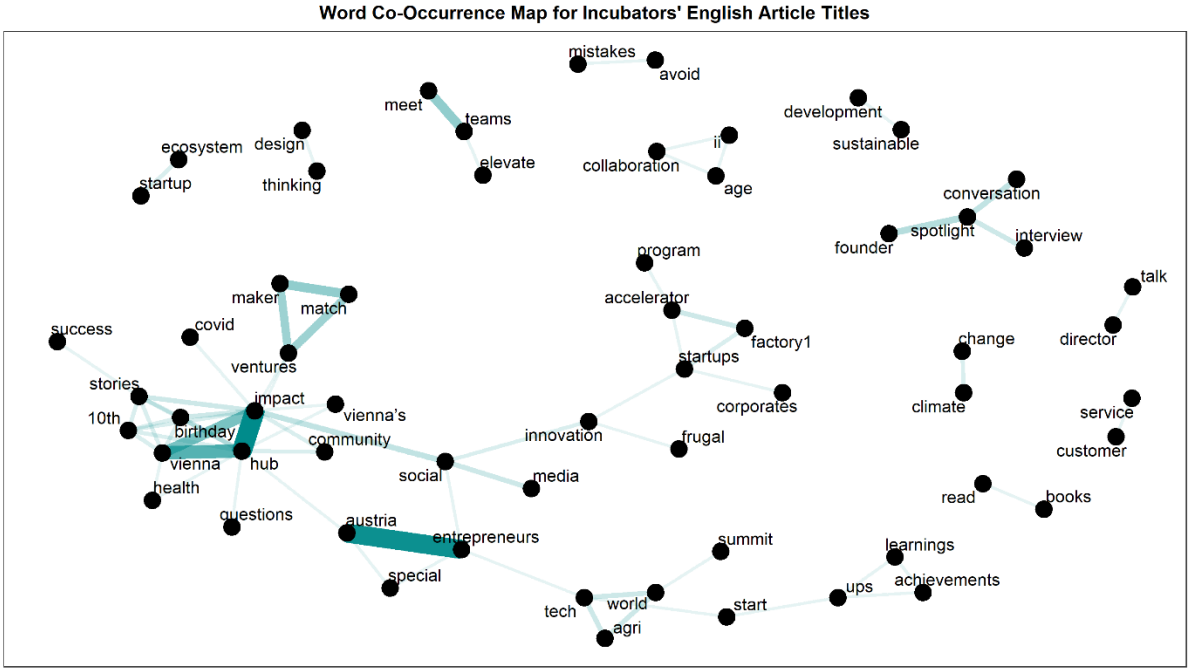

Figure 52

Token Sentiment Analysis

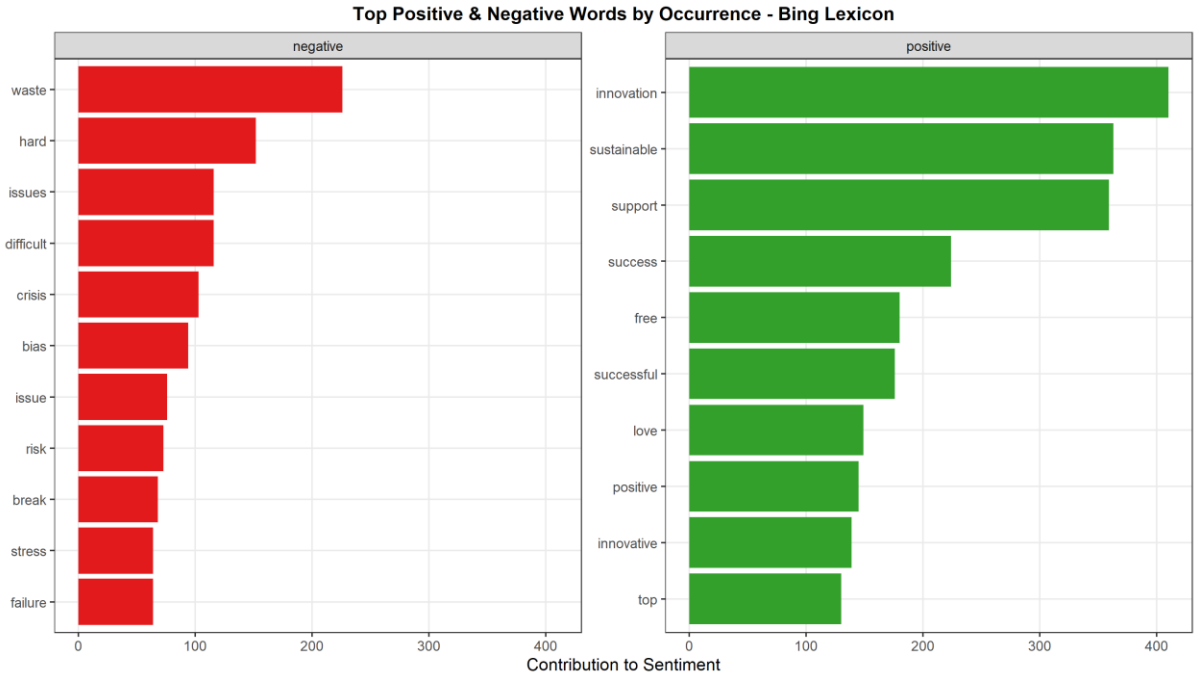

Figure 53

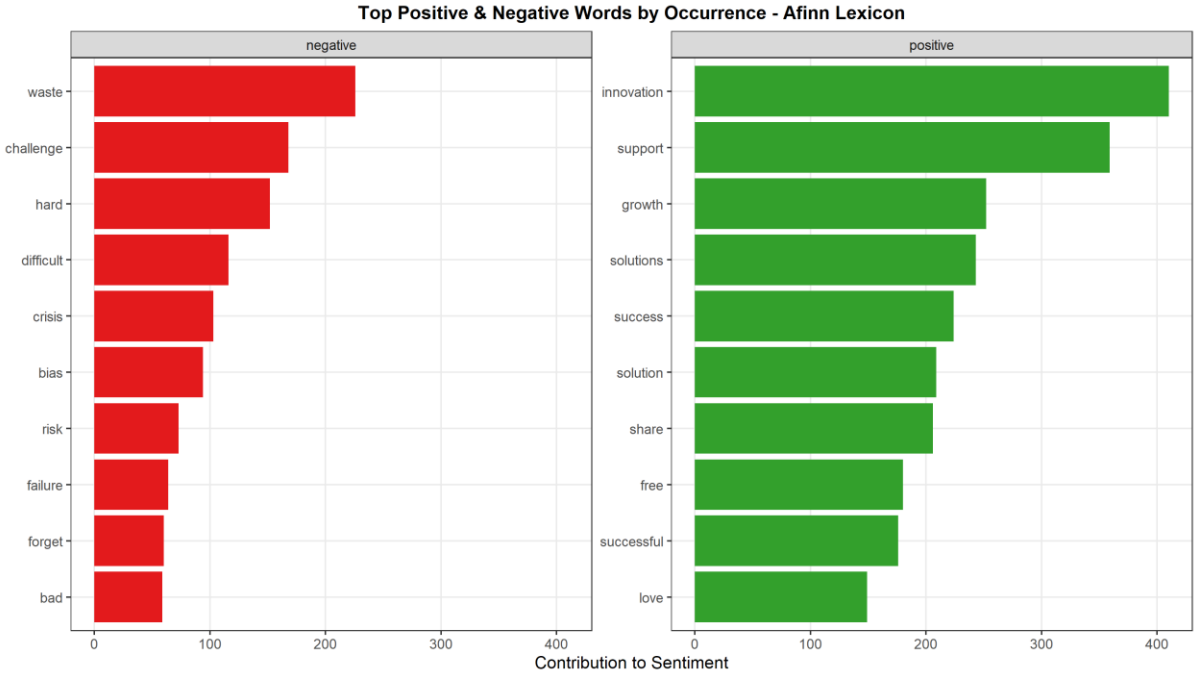

Figure 54

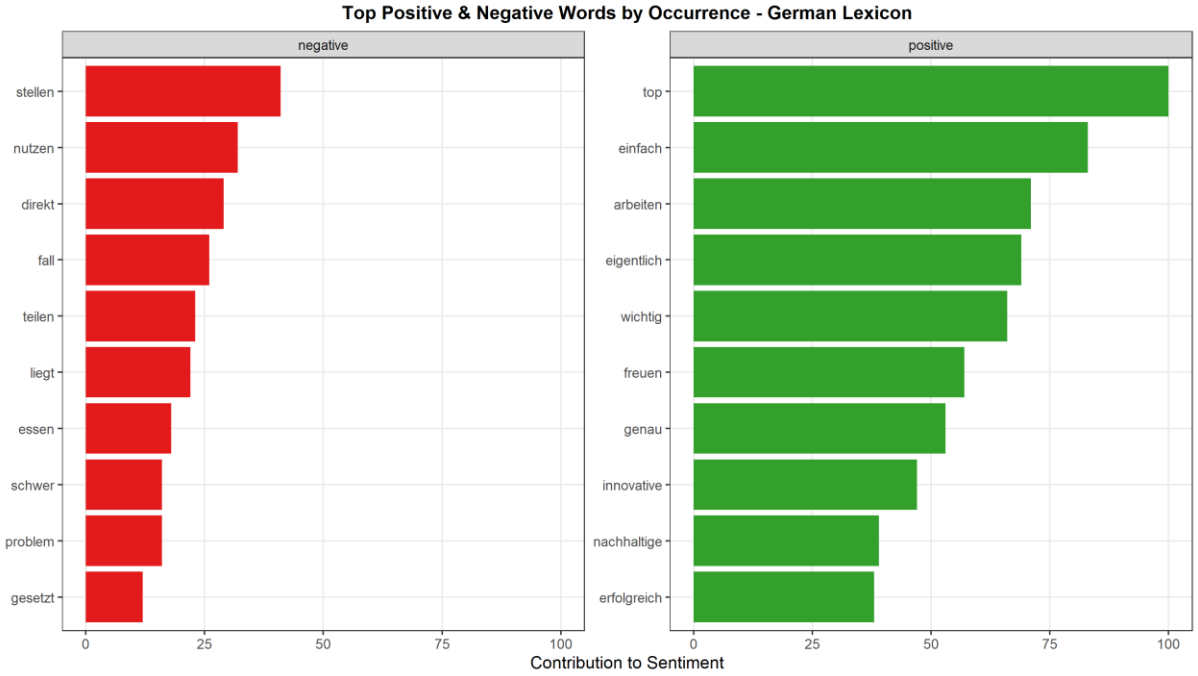

Figure 55

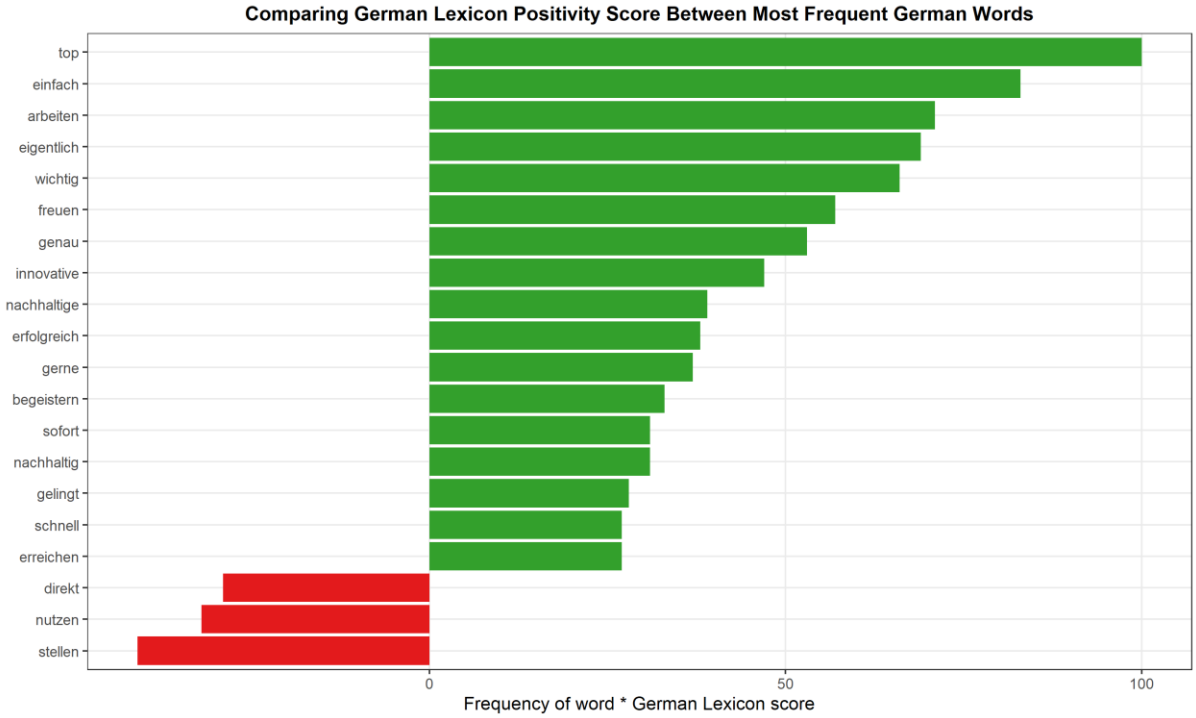

Figure 56

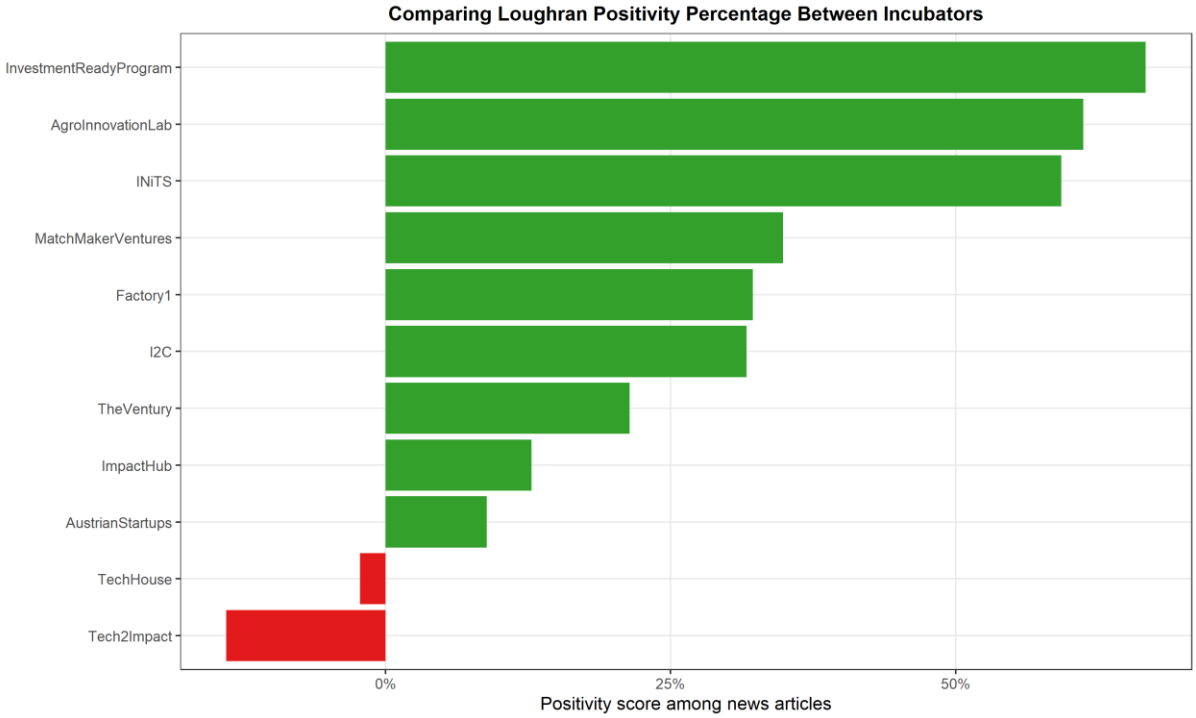

Figure 57