Importing threats and fears? - Using Latent Dirichlet Allocation to disentangle threats and fears on twitter

Daniel Saggau*

May 2020

Abstract

Michael Webb developed a method to examine the impact of technology on occupation matching patent data with job descriptions. Using natural language processing tools, Webb argues that technology challenges the existing standing of different sectors, depending on the task at hand. Moreover, the effects of software, automation and artificial intelligence (AI) deviate. Webb argues that AI will reduce 90:10 wage inequality, but will not affect the top 1%. To complement this analysis, this paper tries to disentangile how the general public perceives threats to employment. Studying 170.000 tweets during the period of the economic lockdown due to the recent COVID-19 outbreak, evidence hints towards the narrative that tweets related to China and migration are loaded with a more diverse set of sentiments relative to tweets related to technology. Moreover, geospatial difference exist between topics. Topic models and a network analysis allow for a glimpse into the general standing of the public opinion. Nevertheless, descriptive data hints towards a bifurcated narrative during the COVID-19 pandemic.

1 Introduction

Opinions about the the future of automation have become increasingly bifurcated, as much as political. The dichotomous nature of technological change have made this topic convoluted. On the one hand, technology is increasing productivity and job demand. On the other hand technology has shaped the nature of work, causing a reallocation of labor supply. Politics has seen a visible movement towards partisanship. The attention of policy makers has been consumed by a number of topics, failing to agree on a common narrative (Autor et al. (2016)). Technological change has been at the center of attention in research on economic growth, as manifested in several pivotal models in economic growth such as the Solow or Ramsey model. While technology is evolving and spreading in an ubiquitous manner, the affect of technology is still largely ambitious as different breakthroughs have led to severely different outcomes. In the last decades, software and robotics have reshaped various industries. Low skill jobs have become redundant with the establishment of industrial robots (Webb (2019)). Irrespective, other technological breakthroughs have challenged the existing standing of jobs,

*Ludwig Maximilian Unversity Munich, daniel.saggau@campus.lmu.de

demanding a higher tier of skills (Autor et al. (2016)). One example is the emerge of software. This has led to a reduction in middle class jobs. Numerous scholars have examined this effect. autor_skill_2003 provided a framework to understand these outcomes. They compare routine and non-routine tasks, suggesting that tasks which are non-routine are less exposed to software. Interpersonal and analytic tasks have been challenging to replace by software according to their findings. Routine manual work on the other hand has been affected the most severely, with routine cognitive tasks following second. Due to the gender gap in the selection of jobs, women have been less effected by technological change relative to men, due their accumulation in professions in fields with stronger interpersonal components. Acemoglu and Restrepo (2019) build a task model to quantify exposure to technology. Webb (2019) builds on this framework, providing a holistic examination of different types of technology. Furthermore, Webb (2019) complements the existing literature by looking at impact of AI and the recent leaps within these fields.

Webb uses natural language processing to provide a holistic understanding of three technological leaps, namely software, robots and artificial intelligence. Webb (2019) provides corroborating evidence that while software and robotics have a negative impact of low skilled jobs, AI is having a negative impact on high skilled labor.

Economic hardship has been part and parcel of this crisis, but attention has been drawn to the fact that this situation may stimulate further investment by large companies. As argued by Autor et al. (2016), during times of economic hardship partisanship and political polarization have been prevalent over the course of history. A study by Autor et al. (2016) suggests that there is descriptive evidence that areas in which there were disruptive economic shocks, voting behavior has changed. But, within public opinion there is no coherent narrative to who or what is responsible for this outcome, nor is there a consistent trajetory for future policy. Looking at recent events, according to the department of labor, approximately 40 million people filed for unemployment in the US over the course of a few weeks ((“U.S. Jobless Claims Pass 40 Million: Live Business Updates” 2020)). The trajectory towards further automation has been amplified due to the recent pandemic. But, who will there be to blame and how will this alter the discurse of politics and inevitably policy making? This study will provide further insights on the specific sentiments in the COVID-19 lockdown.

This paper will use web scrapping to obtain twitter data on convoluted topics. The study will examine data from 19.05.2020 up to the 27.05.2020, using an application programming interface (API) to obtain twitter data. Three methods are used to assess public debate, namely sentiment analysis, a topic model and a network analysis. For the topic model we are using a three-level hierarchical Bayesian model, namely a Latent Dirichlet Allocation with a variational expectation maximization. For reference, the appendix provides a sample outcome using gibbs sampling. For the network analysis, we use ngrams with 3 words. To complement these insights, the appendix contains a brief outlook of tweets on automation and migration. Sentiment analysis is an analysis of feelings, with which will allow one to manifest the general tendency towards certain topics. Moreover, topic modeling and specifically LDA, the method used in this paper, will provide insights into the exact topics suggested in these tweets. These insights may be able to give more precise information on the impact of the lockdown and the rise in unemployment on highly political topics. This may allow us to get a glimpse into

what direction politics and policy making may go.

This paper is structured as follows: Firstly, there is brief introduction about inequality, technology and the role of politics and policy making for the future trajectory of an economy. Subsequently, this paper will summarize the findings and method of the paper ' The Impact of Artificial Intelligence on the labor Market' published by Michael Webb. Further, using twitter data, this paper will illustrate the current standing towards topics relevant for political debate using sentiment analysis. Thereafter, these findings will be compared with the outcome of our topic model and a brief network model with trigrams will be introduced. Lastly, this paper provides a conclusion, following a discussion, connecting recent economic and political research.

1.1 Literature review

The controversial release of Capital in the 21st century by Thomas Piketty (2013), provided a narrative for the dynamics of inequality. While numerous scholars have rejected the central hypothesis in this book, it has had a large outreach. This book was heavily inspired by the general laws of capitalism and historical materialism introduced by Karl Marx, providing a theory of history suggesting that our material conditions will shape our history. Piketty argues that capital accumulation and technology will disrupt the forces of production and economic/political life resulting in a more unequal society (Acemoglu and Robinson (2015)). Acemoglu and Robinson (2015) argue that this theory fails to incorporate the endogenous evolution of technology, accommodating for the fact that technology and its impact are severely dependent on the accompanying institutions and politics. Recent research by Acemoglu and Restrepo (2019) provide an analysis to understand the multifaceted nature of technology. They emphasize the role of the displacement effect and the productivity effect. The displacement effect argues that workers are replaced through automation, because technology occupies their prior positions. Inevitably, this suggest a disentanglement of the relationship between wages and output, with a declining share of labor. The productivity effect claims that due to the reduction in production cost, economic expansion will result in accelerated demand for labor for non-automated tasks. Acemoglu and Restrepo (2019) argue that the productivity effect could emerge in non-automated sectors or the sector which is undergoing automation. Moreover, Acemoglu and Restrepo (2019) state that automation occurs at both the extensive margin, replacing tasks performed by workers, and the intensive margin, replacing tasks performed by machines, deepening automation. They argue that this would also lead to further productivity but countervail the displacement effect. Following this argument, the reinstatement effect, the emergence of novel tasks, take the same direction. Subsequently, Acemoglu and Restrepo (2019) discuss the mismatch between technology and skills. Their theory claims that this mismatch is created by a lack of accommodation of new technologies in the educational system, creating a lag in productivity. The authors provide two further arguments: Firstly, there is crowding out of growth opportunities because of a unidirectional focus on specific technologies. Secondly, the authors argue that excessive automation, caused by misguided US tax incentives has led to the socially disruptive innovation, deteriorating productivity growth. Following this debate, Webb (2019) suggests that the impact of inequality deviates between different technologies. Webb (2019) uses the

static canonical task based model by Acemoglu and Restrepo (2018b).

Webb (2019) argues that he makes three contributions to the field: 1. A general-purpose measure of technology providing a more holistic insight relative to the prior work on specific industries and technologies 2. Impact automation on jobs and wages 3. The impact of AI specifically

Work which relies on interpersonal components, have been less prone to automation and displacement. Due to the gender imbalances in these jobs entailing larger components of interpersonal skills, men are more affected by technology caused displacement. Tasks which entail a larger share of routine, have been more prone to automation and displacement. Especially "muscle tasks", tasks which are connected to muscle work, have been exposed to robots.

2 Summary

Webb (2019) examined three distinct effects, namely the affect of software, robots and AI on employment. The study uses numerous data sets:

- Patent data: Google Patents Public Data (IFI CLAIMS Patent Services)
- Job description data: O*NET database of occupations and tasks (provided by the US Department of Labor)
- Employment/Wage data: Individual level microdata form the US Census 1960-2000 and from the ACS 2000-2018 provided by IPUMS.

Further, exposure is measure to quantify whether a task can be automated by a particular technology. His model was based on a static task based model, originating from Acemoglu and Restrepo (2018b). Emphasis was placed at defining exposure of different occupations. To determine how over the course of time technology has changed, NLP is used capture changes within job descriptions and patents.

2.1 Model environment

The model was introduced by Acemoglu and Restrepo (2018b)- For more information, see this paper.

This section is a recap of the model defined by Webb (2019).

The economy is defined as an entity, entailing one firm. Good X is produced, by combining the product of different occupations O_i. Further here Webb introduced the assumption that elasticity of substitution is constant, manifested in the parameter ρ. While this assumption does simplify the further analysis, inference based on this assumption should be questioned as the author even admits there is evidence that this may not hold in the real world. Hence, let:

$$X = \left(\sum_i \alpha_i O_i^{\rho} \right)^{\frac{1}{\rho}} \tag{1}$$

O is the occupation, while T are the number of tasks.

$$O_i = \left(\sum_i \alpha_j T_{i,j}^{\rho_t} \right)^{\frac{1}{\rho_t}} \tag{2}$$

j indexes the number of tasks. Now we can distinguish between tasks that are exposed to automation versus tasks that are not exposed to automation.

$$T_{i,j} = \begin{cases} H_{i,j} + A_{i,j} R_{i,j} & \text{if automation feasible} \\ H_{i,j} & \text{otherwise} \end{cases} \tag{3}$$

Here R embodies machines while H stands for humans. Further the author uses text data as fundamental input for the exposure scores. The following section provides a comprehensive overview of how these data points were obtained.

2.2 Method:

Natural language processing is used to process text data. Webb (2019) uses the patent data to determine the extent on the given occupation of the given technology. The method is used to extract verb-noun pairs, also called bigrams. Bigrams usually describe word pairs. One can also work with unigrams, but these methods usually involve a substantial loss of information. Ultimately, these pairs are used to quantify overlap between patents and jobs. The method used here is to determine the frequency of different pairs.

Roughtly the preprocessing of the text data can be separated into three parts:

1. Define restrictions and diminishing the dimensionality by tokenizing the data
2. Filtering for stop words and common terms/words
3. Stemming and lemmanizing the terms/words

Frequency is defined as follows:

$$rf_c^t = \frac{f_c^t}{\sum_{c \in C^t} f_c^t} \tag{4}$$

Following the definition by Webb (2019), let: rf_c^t be the aggregate verb noun - pairs relative frequency. c is the specific word-pair and t is the technology. f_c^t be the raw count. C^t be the full set of aggregate word pairs.

This method approximates strongly with word embeddings. Specifically, Webb uses a word parcing algorithm, allow one to obtain information on words within their embedded sentences. Another notorious method is the bag of words-method, in which words are analyzed independent of their structure involving a severe loss of information and reduction in dimensionality. While these bag of words are criticized frequently, they are often common practice. Word parcing methods promise higher accuracy. This method is dominant in computational linguistics but didn't establish as common practice within text data in

economics. This method is recommendable when one has substantial prior information and limited possibilities to split the data into training and test sets for text classification and generative text models. After collecting the data, the words are lemmatised words using a dictionary method, namely the WordNet dictionary. Lemmetisation enables one to group words into a category. This reduces the dimensionality of the data. Subsequently, these word pairs are used for inference. Inference based on text data is also less prominent within research.

2.2.1 Patent data

Firstly, a set of patents are selected, focusing on specific technologies. Webb undertakes selection based on patent titles. Subsequently, verb-noun pairs are extracted from the patent titles. The first step is using a dependency parsing algorithm. This algorithm provides the feature of being able to extract syntactic relationships of words within their sentences. As mentioned by Webb (2019), For each verb, the direct object is attributed. The verbs and nouns are lemmatized. Stop words (e.g. has, use, have) are extracted. The probability of a specific verb-noun pair occurring is calculated.

2.2.2 Job description data

Looking at the job description data, occupations are matched with specific tasks. The dependency parsing algorithm is used. Tasks are ranked by their frequency in the given occupation to provide a weight for these tasks within occupations. Each pair of verb and noun is put into conceptual categories. For that, a hierarchy of concepts is used. Thereafter,the author undertook stemming. Probabilities are measured, generating task and occupation scores. Due to the size of the data set, the paper could not correct for false positives.

WordNet is used to group words into hierarchies of concepts. This is done to maintain conceptual categories which are mutually exclusive at nature. Verb-noun pairs are aggregated and used to match the patent data. The main regression is run using a granular aggregation level, namely WordNet level 3 (and rerun with 2 and 5 for sensitivity), to keep more information. For each exposure score, the weighted average is used to produce overall scores. Weights are based on the O*Net database, which provide frequencies of the different jobs.

To complement the NLP results, an empirical strategy is selected.

$$\text{Exposure}_{,t} = \frac{\sum_{k \in X_i} \left[w_{k,l} \cdot \sum_{c \in S_k} rf_c^t \right]}{\sum_{k \in K_i} \left[w_{k,l} \cdot \| c : c \in S_k \| \right]} \tag{5}$$

define the set of task in an occupation to be K_i. $k \in K_i$ is the task within the set of tasks. S_k is the set, entailing the verb-noun pairs. $w_{k,l}$ is the weight of each task for occupation i. Further specifications can be found in the original paper.

2.3 Results for the specific technology

For each specific technology, the author provides descriptive statistics and an empirical model. The descriptive statistics deals with the impact on occupational wage percentile, exposure by level of education, exposure by percent of female workers in occupation and exposure by age. Moreover, for the empirical strategy, the dependent variable, the change in wages, is 100* change in log wage. Wages are cells mean weekly wage for full time , full year workers in 1980. The variable offshorability is a occupation level measure, inspired by the 2013 paper by Autor and Dorn 'The growth of low skilled service jobs and the polarization of the US labor market'. Moreover, Webb develops an exposure measure to capture the exposure to technology. Data for this empirical model comes form the US Census data base, starting from 1960-2000 and fr0k ACS 2000-2018.

$$\Delta y_{o,i,t} = \alpha_i + \beta \operatorname{Exp}_o + \gamma \mathbf{Z}_o + \epsilon_{o,i,t} \tag{6}$$

The regression model is provided in the references. As the most essential parts of this paper are the effects of technology on different groups, this will be the focus of the summary.

Each section (robots, software, AI) contains information on the least and most effect occupations and their underlying theries.

2.3.1 Robots:

Webb used an employment threshold of 150 to filter out the 5 most and least exposed professions. For reference, the list provided by Webb was included in the appendix. Here, we filter via the number of aggregate word pairs occurrences. Moreover, the 10 most and least exposed professions are provided for all three areas.

index	count	score
Janitors and Cleaners, Except Maids and Housekeeping Cleaners	30	2.813520
Elevator Installers and Repairers	36	2.772875
Locker Room, Coatroom, and Dressing Room Attendants	26	2.758105
Bridge and Lock Tenders	36	2.745791
Farmworkers, Farm, Ranch, and Aquacultural Animals	27	2.745181
Agricultural Equipment Operators	31	2.648833
Ophthalmic Laboratory Technicians	35	2.563559
Locomotive Engineers	24	2.412980
Rail Yard Engineers, Dinkey Operators, and Hostlers	31	2.389656
Helpers–Pipelayers, Plumbers, Pipefitters, and Steamfitters	27	2.357243

As dicussed in the paper, evidence hints towards the reduction of "muscle jobs", and a stronger affect on non-cognitive routine tasks as displayed in the table.

index	count	score
Postal Service Mail Carriers	31	0.0079512
Interpreters and Translators	24	0.0080389
Telemarketers	22	0.0109408
Music Directors	26	0.0133555
Art, Drama, and Music Teachers, Postsecondary	23	0.0166473
Eligibility Interviewers, Government Programs	26	0.0190711
Title Examiners, Abstractors, and Searchers	31	0.0191773
Naturopathic Physicians	24	0.0194494
Foreign Language and Literature Teachers, Postsecondary	21	0.0197163
Correspondence Clerks	33	0.0211703

Occupations with a substantial degree of interpersonal and manual components are less affected by robots. As seen in the table, art related and creative jobs have been amongst the least affected. One see a difference between genders, as women frequently occupy such positions. Hence, one can see that men under the age of 30 are most affected, even more so than women in the same age group having the same educational obtainment.

2.3.2 Software:

One can see that, as mentioned by Webb that here middle and high skilled jobs are exposed strongly. Moreover, one should note that these jobs might have been exposed but necessarily led to the displacement. These affects may manifest differently such as e.g. a reduction in wages.

index	count	score
Agricultural Equipment Operators	31	1.558714
Locomotive Engineers	24	1.483366
Power Plant Operators	43	1.346853
Bridge and Lock Tenders	36	1.332027
Pest Control Workers	28	1.327109
Automotive Engineering Technicians	22	1.307505
Power Distributors and Dispatchers	26	1.295371
Geothermal Technicians	30	1.281601
Police, Fire, and Ambulance Dispatchers	29	1.276018
Surveying Technicians	28	1.260261

Again, cognitive and non-routine tasks have been less exposed to software as suggested by the appearance of creative jobs in the least exposed list. For reference, see the table in the appendix with the selected orginal professions.

index	count	score
Cooks, Restaurant	24	0.0271488
Graduate Teaching Assistants	21	0.0504130
Postal Service Mail Carriers	31	0.0638194
Poets, Lyricists and Creative Writers	24	0.0677374
Social and Human Service Assistants	22	0.0712166
Interpreters and Translators	24	0.0715302
Accountants	25	0.0852841
English Language and Literature Teachers, Postsecondary	25	0.0871249
Demonstrators and Product Promoters	34	0.0896400
Foreign Language and Literature Teachers, Postsecondary	21	0.0929643

2.3.3 AI

Lastly, Webb (2019). examines artificial intelligence. Within the field of AI there are two pivotal pillars. General artificial intelligence deals with AI that targets universal decision making. The second is specific artificial intelligence. This is a type of AI that handles a specific problem, but would not be able to act outside of this framework. One field within specific AI is machine learning and it has become the flagship for artificial intelligence. These field have gained prominence due to fear of automation caused by these technologies. Supervised learning has become highly efficiently, accomplishing better performances than humans in numerous task. Moreover, supervised learning, the counterpart to unsupervised learning, is a method in which a task is defined. Unsupervised learning deals with classification/prediction of of unknown tasks. Here, a task is detected through training. The application of machine learning is accelerating in fields that are targeting the employment of high skilled labor. Prominent examples are the use of machine learning for anomaly detection in medicine and the detection of financial fraud. Mullainathan and Spiess (2017) illustrate in what manner economics has adopted machine learning methods in their paper *"Machine Learning: An Applied Econometric Approach"*. Applications include the use of machine learning to analyse satelite data, allowing one to get insights into poverty data. Especially in the field of economic development this method has gained attention. Further, another application is the prediction of teacher quality in hiring decisions and the analysis of policy success. Below find the table with the most exposed professions:

index	count	score
Civil Engineering Technicians	27	1.457201
Gas Plant Operators	33	1.374805
Locomotive Engineers	24	1.329006
Nuclear Equipment Operation Technicians	31	1.315572
Transportation Vehicle, Equipment and Systems Inspectors, Except Aviation	24	1.300407
Precision Agriculture Technicians	41	1.274694
Nuclear Monitoring Technicians	32	1.238676
Agricultural Equipment Operators	31	1.187799

index	count	score
Critical Care Nurses	33	1.183788
Police, Fire, and Ambulance Dispatchers	29	1.171750

High and low skilled jobs are challenged in this list. AI is deals with the detection of patterns, decision making and optimization. Hence, more advanced occupations are more keen to be effect relative to classical muscle tasks.

Webb (2019) points out that people with undergraduate and graduate degrees from universities are most exposed to AI. Again, we filter for aggregate word pairs occurring more than 20 times. Hence, AI is a threat not only for medium and low skilled jobs. Therefore, this should alert policy makers when accounting for these factors. As mentioned above, this exposure may embody in a reduction in wage or a higher skill set. Numerous authors such as Frank et al. (2019) argue that this exposure is leading to higher entry requirements and technical skills for jobs. Frank et al. (2019) mention the mismatch between what college education is offering and what the future market is requiring.

Overall, the relationship between technological automation and the labor market is considered non-linear. As one could see in the research, the results were different for different groups. One should note that little research was done on sectors or occupations which have benefited from further automation. The study solely focused on the negative repercussions from further automation, making predictions based on numerous assumptions and simplifications. One of the post pivotal simplifications is the fact that elasticity of substitution is assumed to be constant at various levels. Moreover, as mentioned in the discussion section of the paper, ownership of capital may benefit from automation. The largest concern expressed by the author was with respect to timing. The affect of technology on the labor market lags compared to the time the patent is instantiated. Thus, what should have been done to optimally capture the affect of technology and labor is determining a optimal time lag for the relationship.

index	count	score
Biological Science Teachers, Postsecondary	22	0.0000000
Art, Drama, and Music Teachers, Postsecondary	23	0.0000000
English Language and Literature Teachers, Postsecondary	25	0.0000000
Cooks, Restaurant	24	0.0000000
Morticians, Undertakers, and Funeral Directors	22	0.0000000
History Teachers, Postsecondary	21	0.0114706
Postal Service Clerks	31	0.0124939
Nonfarm Animal Caretakers	28	0.0148294
Graduate Teaching Assistants	21	0.0150491
Telemarketers	22	0.0159202

The paper concludes that while robots and software target jobs requiring low skilled labor,

artificial intelligence is tackling the jobs of high skilled workers. Hence, this study provides evidence that technological change will impact the work force for the majority of the work force and not only the lower skilled sectors, prone to replacement. One should note here that these results need to be considered with caution. On the one hand, the method that was used here was relatively reliable compared to other methods not using word embeddings. Nevertheless, other algorithms are not compared or benchmarked within this study. Mixed method models could provide a more comprehensive understanding of the labor market. Moreover, are these results mostly descriptive and not allowing for causal inference or interpretation. As pointed out by Autor et al. (2016), numerous effects lead to change in employment. Factors such as economic shocks or trade with not included in the regression model. Factors such as institions or political frameworks are difficult to include in a model. Irrespective, as pointed out by Acemoglu and Robinson (2015), these factors are pivotal for the future of growth. Especially the inequality measure within the paper does not account for the fact that the impact of AI depends on the pace at which these technological are implement and the quality of the surrounding policies. Hence, while it is indicative, it definitely should not be interpreted as sound prediction of the impact of AI. A further discussion on the implications follows in the dicussion.

3 Analysing threats and fears

As mentioned above, our institutional framework and political realm will determine how a society accommodates changes in employment. As suggested by Acemoglu and Robinson (2015), inequality and economic trajectories are dependent on the institutions and policies they are embedded in. Hence, it is important to understand fears and threats within the general public. Autor et al. (2016) undertook a study examining how trade exposure with china has changed voting behavior. Their study suggests that the US has further polarized, seeing stronger movement by predominately white non-Hispanic males moving towards the political right and minority dominated districts shifting to the political left. As stated in their research, the polarity of ideological beliefs about the source of economic challenges and their political responses corroborate theories about in-group/out-group identification. Rather than disentangling these arguments to polarize at a common narrative, the political sphere is moving towards a competition of group-centric resource allocation (Autor et al. (2016)) Further, evidence indicates that trade has potentially been one of the driving factors for changing voting behavior because of the spatial density of blue color labor that has been challenged by trade agreements with China. Technology on the other hand has altered the existing standing of both blue and white color labor. The authors suggest that technology in combination with other forces, has also contributed to the decline of blue color labor in these concentrated demographic and geographic areas. While Autor et al. (2016) admit that evidence is merely indicative, they argue that economic and political polarization and therefore the rise of partisanship may be caused by disruptive economic shocks partially caused by trade with China. This effect is according to their study amplified due the manifestation of these disruptive shocks in certain geo-spatial areas, accelerating "vehemence at the polls" (Autor et al. (2016)).

Trade, international relations, technology and migration are pivotal pillars of the narrative in

the political realm and shape policies. This paper seeks to illustrate the recent sentiments towards work, using twitter data. Ultimately, the sentiments of a society will shape the path of future voting behavior and therefore the degree of partisanship within an society.

The section of this paper is structured as follows:

3.1 Text data in Economics

This method will provide a brief introduction into the different methods for text data usage in economics.

3.1.1 Document decomposition

The bag of words-method is a frequently used approach to representing data (Grentzkow et al., 2019). The words within a document and decoupled from structure and order, leaving one with phrases depending on the number of words specified. N-grams are phrases with n number of words. Social science research frequently uses unigrams, N-grams with only one word, because of the computational cost of using multiple words. Here one has to weight dichotomous conundrum of using more words, allowing for more precision, opposed to the additional time spend. As argued by Grentzkow et al. (2019), the recommended approach is to start with a unigram, and then depending on the outcome, and then to evaluate whether a more fine-grained approach is necessary.

3.1.2 Methods for count and attribute analysis

Grentzkow et al. (2019), divide methods for count (ci) and attribute (vi) into four groups, namely (1) dictionary-based methods, (2) text regression methods, (3) generative language models, and (4) word embedding methods. Further methods exist, also entailing mixed methods of these suggested methods, providing promising results. Numerous methods emerged for different purposes and with a varying degree of complexity. This paper paper will use dictionary methods and generative text models. Figure 1 provides a brief overview, inspired by the classification used by Grentzkow et al. (2019).

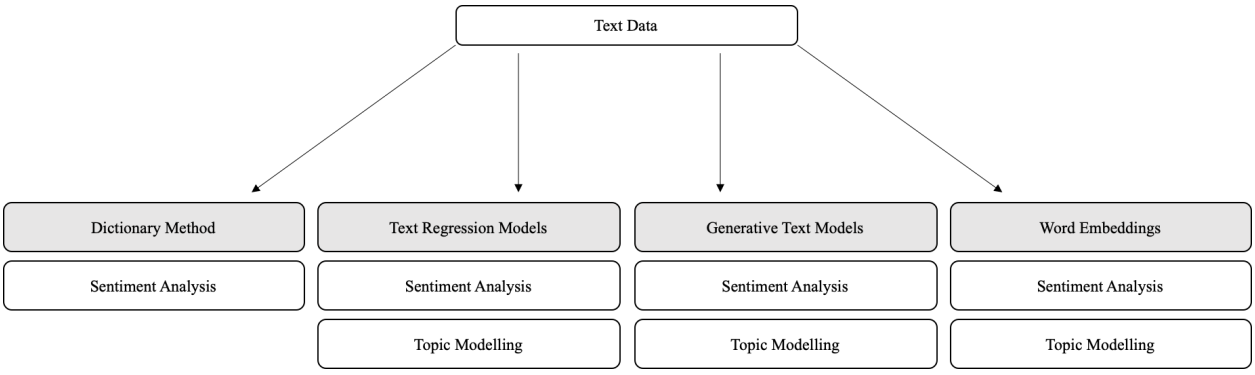

Figure 1: Overview of count and attribute methods

3.1.2.1 Dictionary Method Dictionary methods, using a lexicon, are recommended when prior knowledge is substantial, and there are reduced opportunities to appropriately separate data into training and test data. This method is used predominately for sentiment analysis, and has gained prominence within social sciences. They do not involve statistical inference, as one solely specifies the estimated value of vi as a function of ci (Gentzkow, Kelly, and Taddy (2019)). Dictionaries e.g. "nrc", provide predefined dimensions ("anger", "fear", etc.). Compared to other methods, this method has the tendency to provide a wide term coverage. Irrespective, due to the finite number of words within a lexicon and a finite number of predefined sentiments, dictionary methods are not always recommendable (Ferri, D'Andrea, and Grifoni (2017),). Gentzkow, Kelly, and Taddy (2019) argue that in circumstances in which prior knowledge is substantial and there are limited possibilities to split the data intro training and test set, dictionary methods are the preferred tool of choice.

Specifically, "nrc", "afinn" and "bing" are the three lexicons used here. "nrc" allows for the highest degree of specification when it comes to sentiments. It provides insights into "fear", "joy", "positive", "negative", "trust", "disgust", "anger", "anticipation", "surprise", and "sadness". "afinn" examines integer values ranging from 5 to -5. "bing" distinguishes between binary positive and negative categories (Silge and Robinson (2017))

While results are promising, Gentzkow, Kelly, and Taddy (2019) argue that other methods could potentially outperform dictionary methods, providing a higher degree of accuracy. One of these methods is the text regression method.

3.1.2.2 Text regression Method The text regression method is similar to the generative language model, but both differ in their starting point. Text regression methods starts by using the conditional expectation of the attribute vi. Here one uses the count to examine the topic. With generative text models, one would reverse this order, examing the problem from a causality like perspective. Note that notation needs to be adopted properly. The data is split into training and test set, and then the learner regresses the prediction of the training set onto the test set.
Machine learning algorithms have been one of the dominant tools within this branch of text analysis (Gentzkow, Kelly, and Taddy (2019)). Also algorithms such as random forests have been promising. But, when not being able to define the task unfront, other models such as generative text model allow for unsupervised machine learning in which no specific trajectory is defined.

3.1.2.3 Generative text model This method is currently used frequently for topic modeling. Latent Dirichlet Allocation (LDA) is a generative probabilistic method within the field of topic modeling for text mining (Blei (2003)). LDA is an unsupervised machine learning method, hence a method used for prediction of an un-specified pattern (Gentzkow, Kelly, and Taddy (2019)). LDA specifically is a bayesian hierarchical model. Compared to hard clustering, words can embedded in a multitude of our latent factor, in this case multiple topics.

The figure, which was originally provided by Blei (2003) illustrates the process, providing further visual evidence why LDA is considered a hierarchical model. This is the smoothed

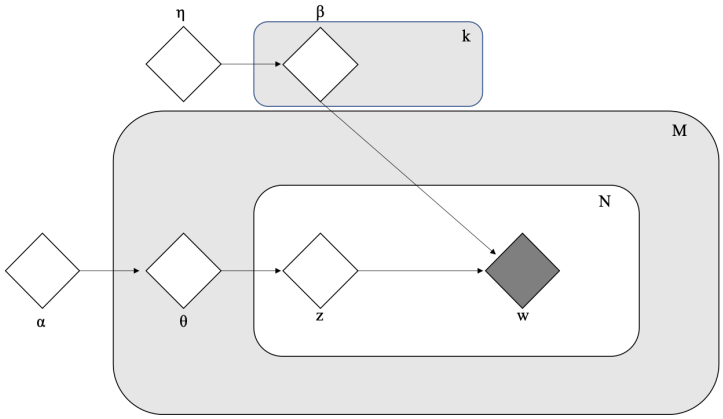

Figure 2: Graphical illustration of smoothed LDA

version of the LDA. Here, alpha is a k vector. Theta is the proportion of the topic distribution. Beta is the term distribution replicate. z is the latent factor, namely the topic. w is the word. N is the set of topics. M is the document in which these objects are embedded in. In the regular model, k and êta are not included. The reason these factors are included is that in large document corpora, some words appear infrequently. A corpus is a collection of M documents. Maximum likelihood estimates would categories these words a 0 probability. Hence, to accommodate for that, all words are assigned a positive probability.

The three-layered structure differentiates this model from other methods. In the first layer, the corpus level, we have alpha and beta. In the second layer, we examine theta, being the document level. In the third layer, we examine z and w, at the topic level.

Cluster methods such as the Dirichlet-multinominal cluster method entail two layers namely the once sampled for the corpus and once for the word level. Given that in this research the package "topicmodels" is used, the model specification and mathematical derivation follows the specification given Grün and Hornik (2011) and the original introduction of the model given by Blei (2003). Following the elucidation by Grün and Hornik (2011), in the first step we need to define the term distribution β. This indicator enables insights into the likelihood of a term occurring in a topic:

$$\beta \sim Dirichlet(\delta) \tag{7}$$

Subsequently, we define the proportion of a topics for our data. Here this parameter is defined as θ and defined as suggested by Grün and Hornik (2011) as follows:

$$\theta \sim Dirichlet(\alpha) \tag{8}$$

Further, given that we assume conditional independence between the tokenized words, we need to specificy the distribution to be multinominal distribution for the.

$$z_i \sim \text{MN}(\theta) \tag{9}$$

The package allows for LDAs with both Variation expectation maximization algorithms (VEM) and Gibbs sampling.

Variation expectation maximization algorithm (VEM):

The expectation maximization algorithm entails two steps. Firstly, the values of the variational parameters are optimized to define a posterior probability. Secondly, using these posterior probabilities, we look at the lower bound of the log likelihood. We are interested in the parameters alpha and beta. Specifically, we want to define the maximum likelihood estimates for each document, entailing sufficient statistics and accounting for the prior defined posterior probability. As illustrated by Blei (2003)," steps are completed until the lower bound on the log likelihood converges.

$$\begin{aligned} \ell(\alpha, \beta) &= \log(p(w|\alpha, \beta)) \\ &= \log \int \left\{ \sum_z \left[\prod_{i=1}^{N} p\left(w_i|z_i, \beta\right) p\left(z_i|\theta\right) \right] \right\} p(\theta|\alpha) d\theta \end{aligned} \tag{10}$$

As pointed out by Gentzkow, Kelly, and Taddy (2019), topic models use a variational inference. This variational inference, allows us to get information on the lower bound of the log likelihood, which is needed for the parameter estimation. This method is used to ensure proximity to a certain parametric family, allowing for approximation of the posterior probability. In bayesian statistics, inference is estimated using a posterior probability, a probability accounting for and embedding prior knowledge. Variation inference is a tool to accommodate the complexity of challenging posterior probabilities. As pointed out, variation inference competes with other methods such as Markov chain Monte Carlo sampling. As pointed out by Blei, Kucukelbir, and McAuliffe (2017): *"Variational inference tends to be faster and easier to scale to large data—it has been applied to problems such as large-scale document analysis, computational neuroscience, and computer vision. But variational inference has been studied less rigorously than Markov chain Monte Carlo sampling, and its statistical properties are less well understood."*.

This method was introduced by Blei (2003). One of these approaches is the Kullback-Leibler divergence. As a reminder: Alpha is a k vector. Theta is the proportion of the topic distribution. Beta is the term distribution replicate. z is the latent factor, namely the topic. w is the word.

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} \text{D}_{\text{KL}}(q(\theta, z|\gamma, \phi)\|p(\theta, z|w, \alpha, \beta)) \tag{11}$$

The variation distribution is defined as follows:

$$q(\theta, z|\gamma, \phi) = q_1(\theta|\gamma) \prod_{i=1}^{N} q_2\left(z_i|\phi_i\right) \tag{12}$$

This results in the following specification for the variational parameter in a lower bound:

$$\log p(w|\alpha,\beta) = L(\gamma,\phi;\alpha,\beta) + \mathrm{D}_{\mathrm{KL}}(q(\theta,z|\gamma,\phi)\|p(\theta,z|w,\alpha,\beta)) \tag{13}$$

Subsequently, we can define the lower bound $L(\gamma,\phi;\alpha,\beta)$.

$$L(\gamma,\phi;\alpha,\beta) = \mathrm{E}_q[\log p(\theta,z,w|\alpha,\beta)] - \mathrm{E}_q[\log q(\theta,z)] \tag{14}$$

This allows us to minimize the difference between our subjective posterior probability and the true posterior, approximating the best possible fit for our posterior. Steps are repeated until the lower bound of the log-likelihood converges Grün and Hornik (2011).

3.1.2.4 Word Embeddings As mentioned above, this method is predominantely used in computational linguistics. Here, words are kept in their orignal structure. These word embeddings, require a higher degree of dimensionality, as we are not reducing words to their mere appearance but account for structures. While these methods have been prosiming as suggested in recent research, they are also computationally costly. Given the prior knowledge and the data these factors need to be balanced.

3.2 Data

This paper examines the sentiments of different threats to employment. Specifically, this paper studies twitter data, using a data set of 170.000 twitters. Following keywords are used: “migration”, “china”,“employment” and “automation”. For each topic, 30.000 tweets are exacted, using keywords except for china and employment for which 40.000 tweets were collected. Subsequently, the tweets are cleaner. Keywords, numbers, websites, special signs, symbols, selected words such as the actual word itself and white spaces are removed. Stemming is used to derive the stem of the words. One does not lemmatise but this method could be use to aggregate results and reduce dimensionality in future studies. Further, for each topic sentiments, frequency, and polarization are analyzed. Additionally, Latent Dirichlet Allocation (LDA) and text classifiers for each topic are used. Last but not least, we provide a basic illustration of how to create a twitter bot algorithm. Future research could examine topic specific bots.

3.3 Sentiments

This section will examine the different topics and their respective sentiments expressed on social media. Here, we are using a dictionary method.

3.4 Sentiments towards China

China has become a salient topic within the media due to the corona outbreak. Trade relationships have been convoluted due to political movements in the united states. This

section will provide insights into the different sentiments towards China. Three different dictionaries are compared. One can see that the outcome deviates substantially.

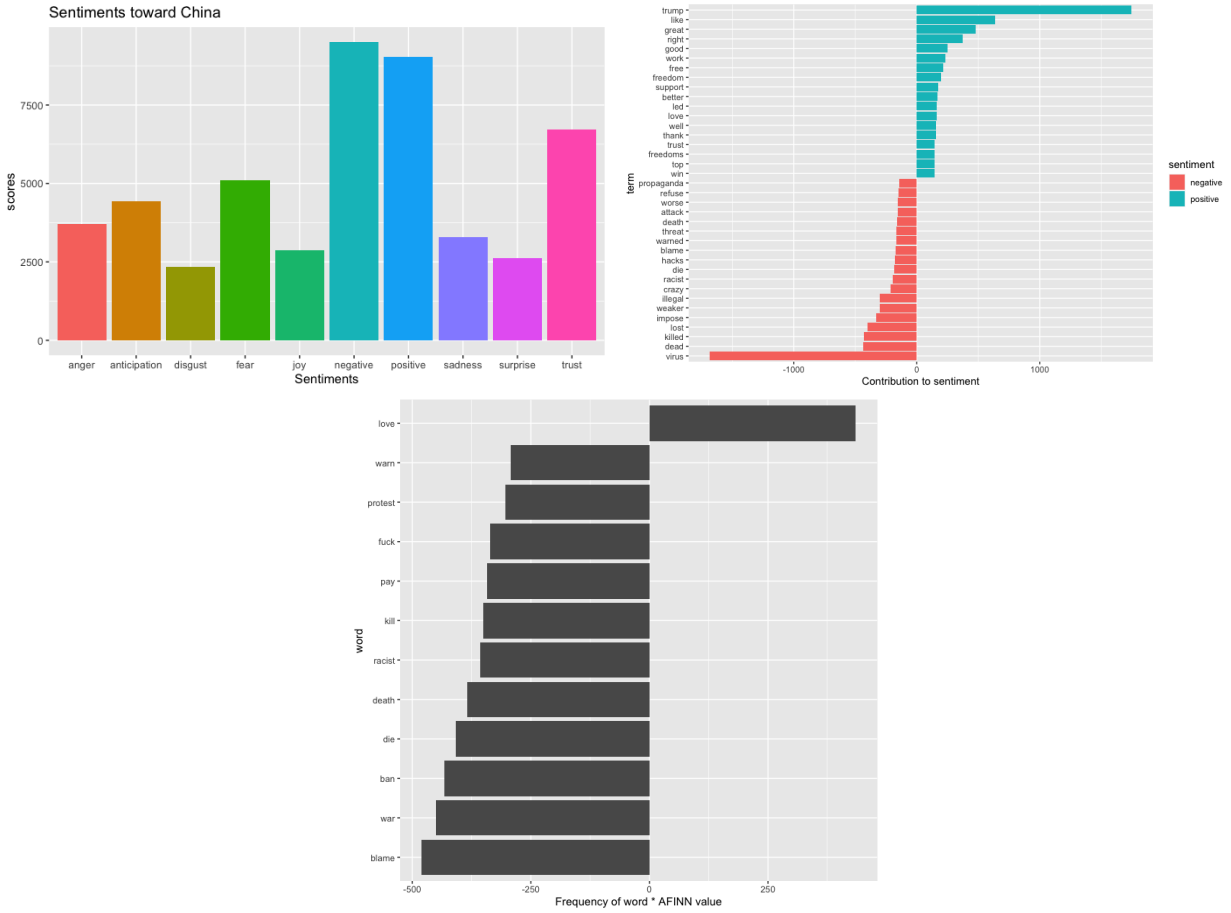

Figure 3: Sentiments towards China

Looking at the "nrc"plot, the overall sentiment score is the lowest. This is due to the fact that the overall score compares both positive and negative feelings. As the "nrc" dictionary is one of the broadest, words that might fall into a simple positive or negative category as done in the "bing" method can be displayed in a nuanced manner. The "bing" method ranks second in overall score. Due to the lack of further categories, it is challenging to get further insights into in which context these words were used. Especially the "afinn" method map a very much stronger negative sentiment stance towards china. Hence, these results should be interpreted with caution. The "afinn" method has provided us with the a detailed glimbse into the most frequently used words, love being the only positive word. This method also tends towards a severely negative narrative.

Ultimately, these results suggest that the dictionary method with unigrams is unstable. There is reason to believe that the extend towards certain sentiments depends heavily on the method used. Irrespective, all these methods also tend towards a negative narrative.

3.5 Sentiments towards employment

This section analyses the sentiments during the lockdown with respect to employment. The keyword 'employment' was used.

Again the three dictionary methods are compared. The "nrc"method suggests that positive, negative, trust and anticipation are most frequently used sentiments. Compared to the China results, the overall score is positive, with positive sentiments outperforming negative sentiments. When examining the "bing" method, one can seee that positive words outperform the number of negative words in frequency. The overall picture is coherent with the "nrc"method. The "afinn" scores are again more balanced compared to the outcome with China (Silge and Robinson (2017)) There seems to be a more nuanced sentiment stance towards this issue, when focusing on this method.

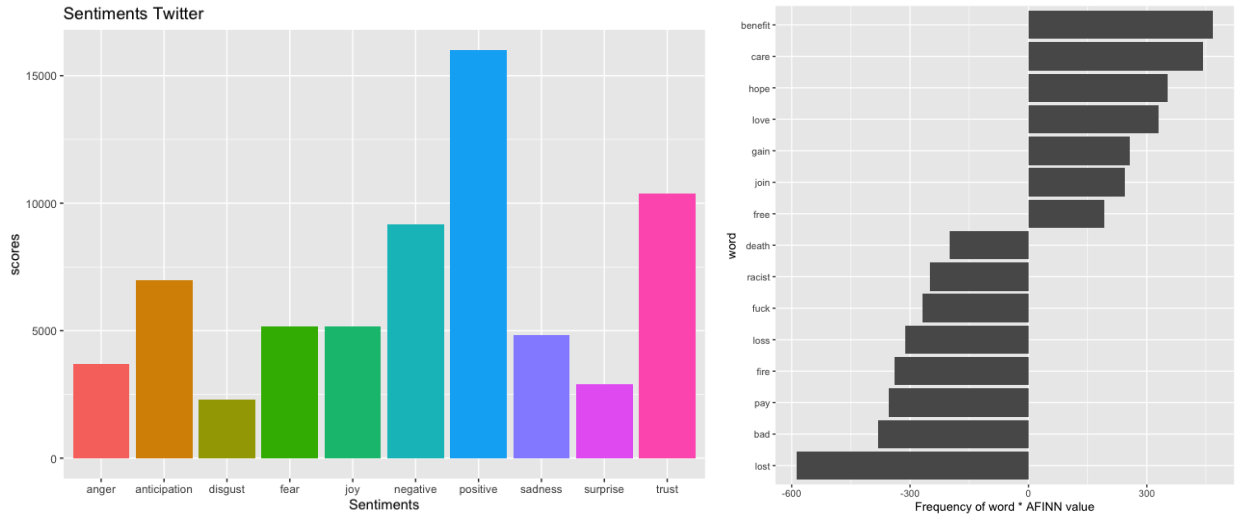

Figure 4: Sentiments towards employment

Ultimately, the sentiment analysis for employment suggests a more nuanced stance.

To get more specific insights, one needs to examine more concise methods such as topic modelling.

3.6 Sentiments Automation

This section illustrates sentiments towards automations. For this section, we used 30.000 tweets in the time frame between the 19.05.2020 and the 27.05.2020. First we compare different dictionary based methods. To examine where people talk about automation, a map was creaeted with the location. Unfortunately, the geospaital data in the tweets was limited. Irrespective, the data illustrates that automation tweets accure for instance in the US more frequently in the country side and the southern states. The graph can be found in the appendix. Further looking at the sentiments towards automation, one can see sizable differences between dictionary methods.

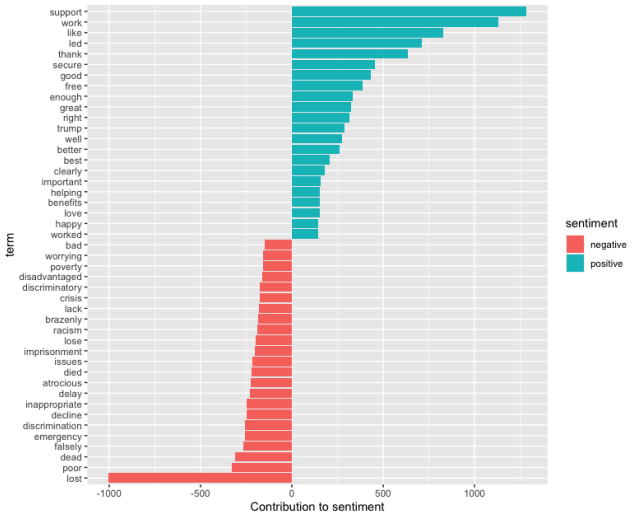

Figure 5: Sentiments towards Employment

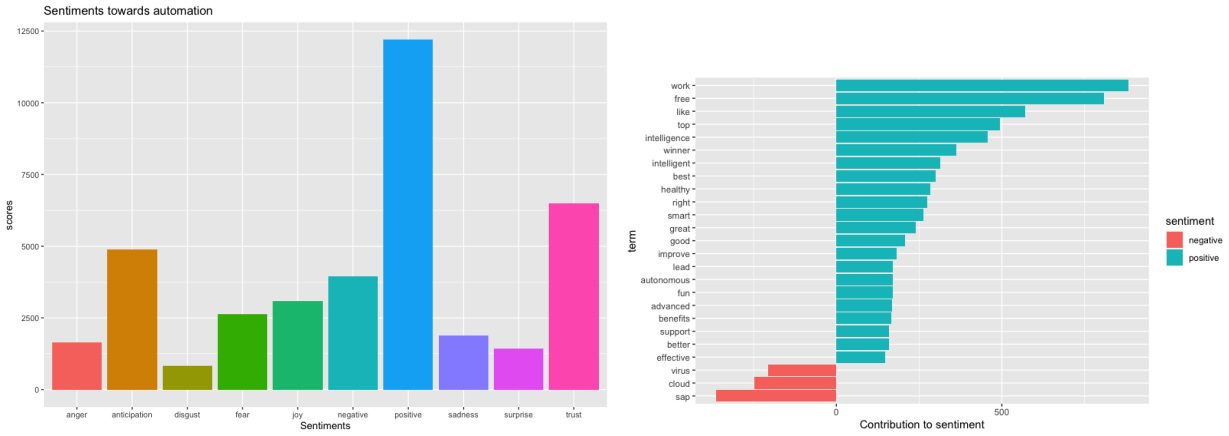

Figure 6: Sentiments towards Automation

Due to the lack of negative terms in the "bing" chart, we provide a second chart showing contribitions most frequently made per sentiment.

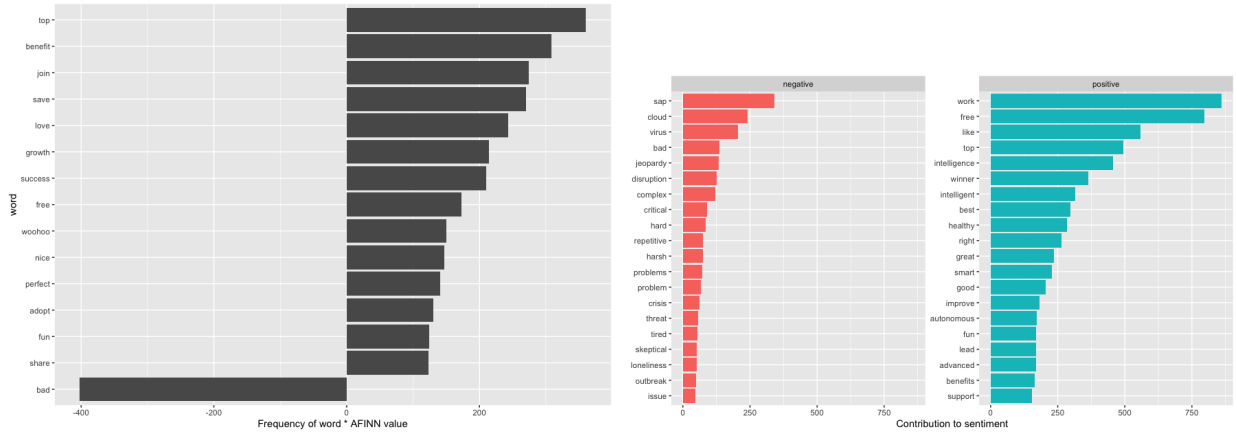

Figure 7: Afinn and bing method entailing contribution per sentiment

3.7 Sentiments Migrants

This section illustrates sentiments towards migration. Examining where these tweets were posted, one can see that the borders were more likely to post tweets on the countries at the border. Moreover, same holds for other countries tweeting about migration. This suggests that these issues are talked about more frequently in areas where migration in a visible issue, while other geographic areas that are landlocked and far from borders don't put this issue at the center of attention. The graph is attached in the appendix. This section uses 30.000 tweets.

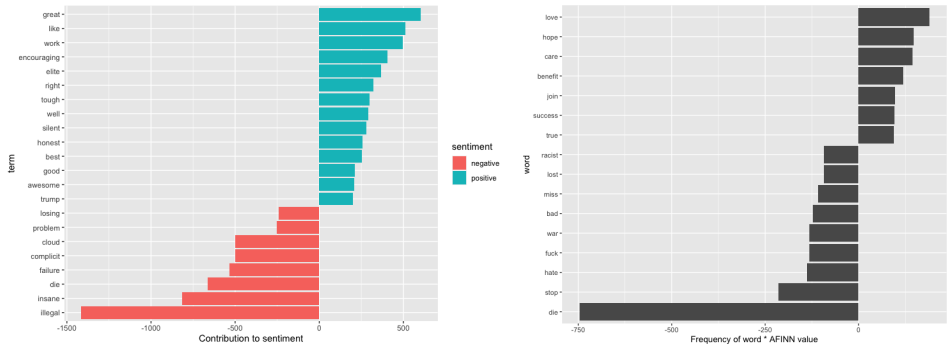

Figure 8: Sentiments towards Migration

Looking at the "afinn" chart, we can see that negative term frequencies outweight positive ones. This is a different results compared to the outcome given by the "nrc"chart. Again, one should note that these unigram method do not account for the embedded context such as "no fear" which would be look at as "fear" and "no" separately.

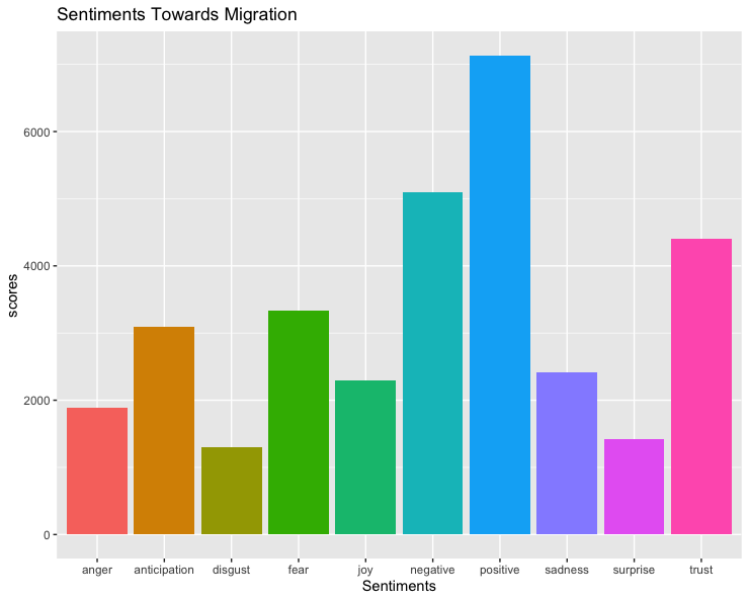

Figure 9: Sentiments towards Migration

3.8 Topic Modelling

Compared to unigram and ngrams, topic models are mixed-membership models. Grün and Hornik (2011) further elucidate that unigrams and ngrams each word is drawn from the distribution of the topic. Topic models allow words to be ascosciated to multiple topics. Lochard Ditricht Allocation is a bayesian mixture model, applicable to data for where topics are uncorrelated (Grün and Hornik (2011)). Mixture models account for the exchangability of words and documents (Blei (2003)). Further Blei (2003) point out that exchangibility does not imply indepedent and identically distributed but rather suggests a condional independence and indentical distribution. For this research the "topicmodels" is used, using the VEM algorithm.. First we use our corpus and summarize the rows, creating a document-term matrix with the word frequency. Then, all the documents with the frequency 0 are removed.

Next, we define

As the name suggests, this model loosens this assumption of having uncorrelated topics, and allowd for topics to be correlated as argued by Grün and Hornik (2011). On a brief note, alpha should approximately be 50/k, k being the number of topics. We define k to be 5. Hence we define alpha as 10.

In the subsequent section one will discuss the topic groups for the four different terms.

3.8.1 Employment

Looking at the employment chart, we can see a debate about how to find different solutions. Due to the recent unemployment, there are more posts related to finding work. Moreover, it appears that numerous post focus on law and policy. Despite filtering for the words "employment" and "job", the word stem employ is depicted in the chart. This illustrates

that the filtering method and cleaning method could have been improved. Especially topic 4 seems to focus on help by the state. Generally it appears that little discussion for this keyword is focused on blaming groups. Looking at the second chart focusing on China, one can see a different narrative.

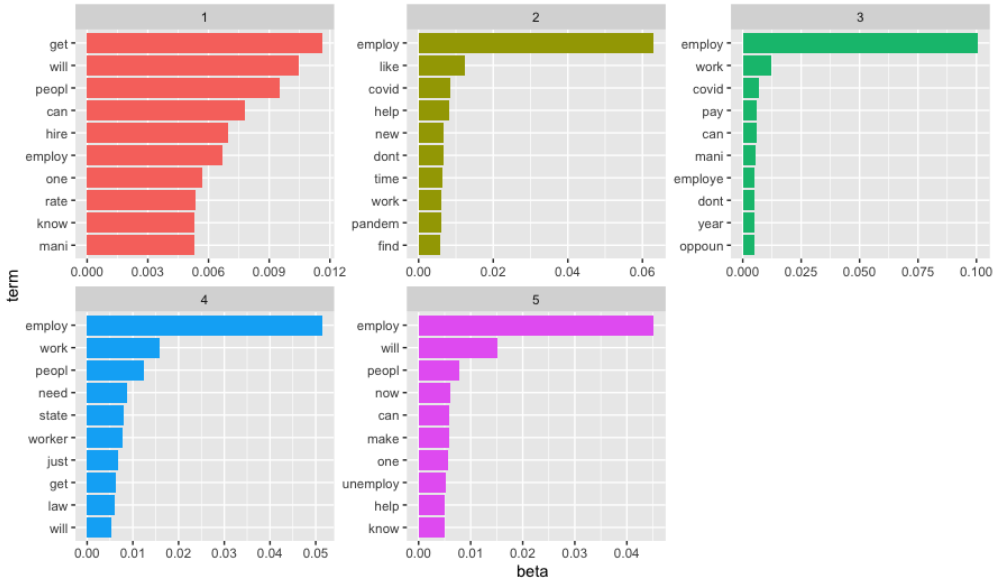

Figure 10: LDA for employment

3.8.2 China

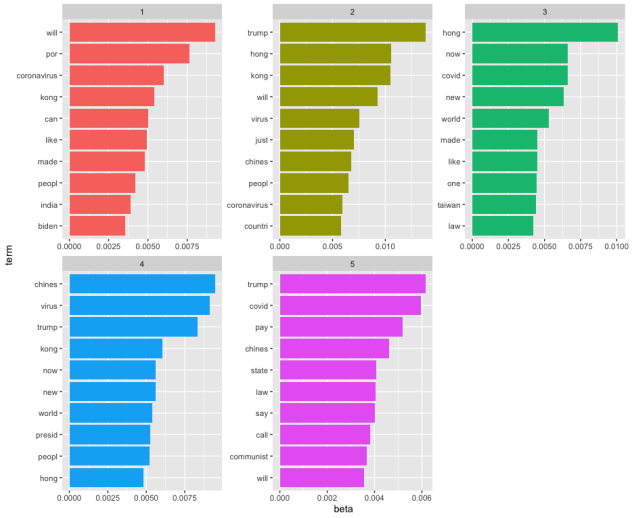

Figure 11: LDA for China

Numerous topics center around Hong Kong, either depicted as “hong” or “kong” or both appearing as seen in topic 1,2,3 and 4. Further, numerous tweets are related to president Donald Trump as seen in topic 2,4 and 5. Also the term people appears numerous times.

But given that we dont know exactly in what context to which other terms, we would need insights into bigrams and trigrams to see in what context people was used. It could have been peoples republic of china, people of Hong Hong/Taiwan or people in context to America. Employment and China dont seem to be the focus of the 5 topics in this chart. To get more information on automation, lets look at the LDA for this term.

3.8.3 Automation

Looking at automation, one can see that the center of attention is on how to use technology and automation to improve the future of work. None of the topics appear to entail hateful sentiments. Verbs such as use,can and learn appear frequently.

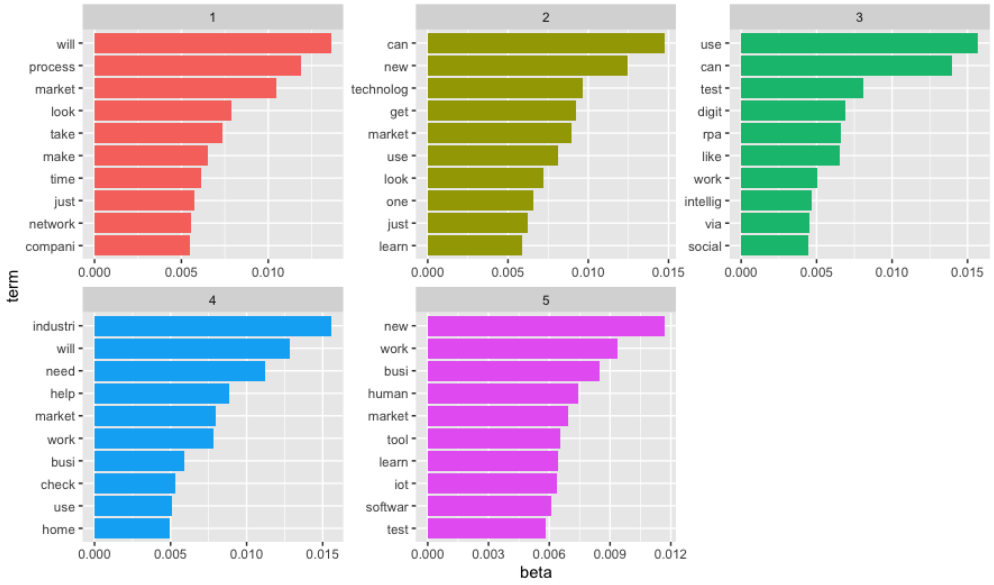

Figure 12: LDA for automation

Generally speaking, these results hint towards a future oriented outlook rather than a narrative of fear and threat. In light of the excess automation argument given by acemoglu_automation_2019, this may be something of interest to study in the long run. Especially given the recent debate about the future of work and an amplification of automation due to the crisis and the exposed vulnerability of human workers. This outcome corroborates the argument given by Acemoglu and Autor that there are overly positive sentiments towards technology relative to other threats. Lastly, lets examine another frequently discussed issue.

3.8.4 Migration

While evidence is mixed, some topics suggest that the advantange of migrants is considered. Illegal immgrant appreard both in topic 1 and topic 5.

For reference, the gibbs method outcome for one sample is included in the appendix.

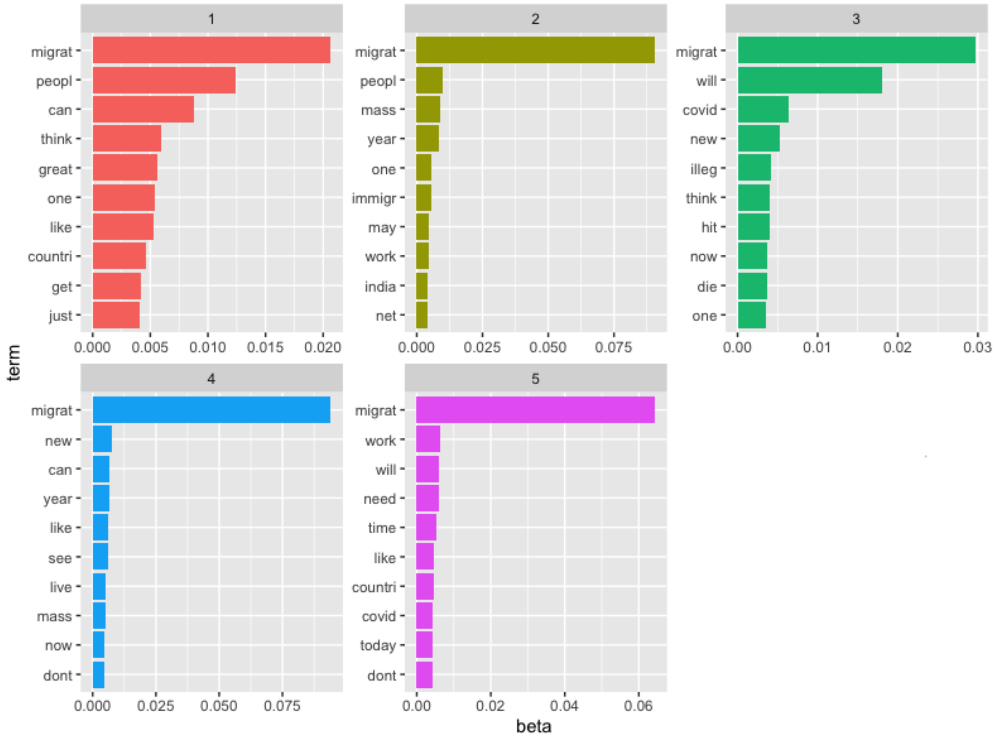

Figure 13: LDA for migration

3.9 Networks

To further visualize common topic debates in twitter data, we use a network graph. Network graphs are used by creating ngrams, examinig words that frequently appear together. Both bi- and trigrams were used. Only the trigrams are reported as their performance was superior. The model was built as follows: Using the clean corpus which was already filtered for stopwords, punctuation, websites, usernames, emojis and other symbols, and filtering for specific terms related to the search term, we build a tidytext document. Then, using this data set, we tokenize the words. For the trigrams we use ngrams with n being equal to 3. Words that appear less than 20 times were removed.

3.9.1 Network: Automation

This visualization suggests that there are 8 to 9 major topics. Not all of them are related to automation directly, as suggested by the one node with twitter billionare jack dorsey.

Numerous tweets wer focusing on machine learning and artificial intelligence. The benefit of this method is that we can see relationships without defining a set number of topics. Numerous of these nodes focus on how to utilize automation during the crisis. Moreover, we will look at the network for migration.

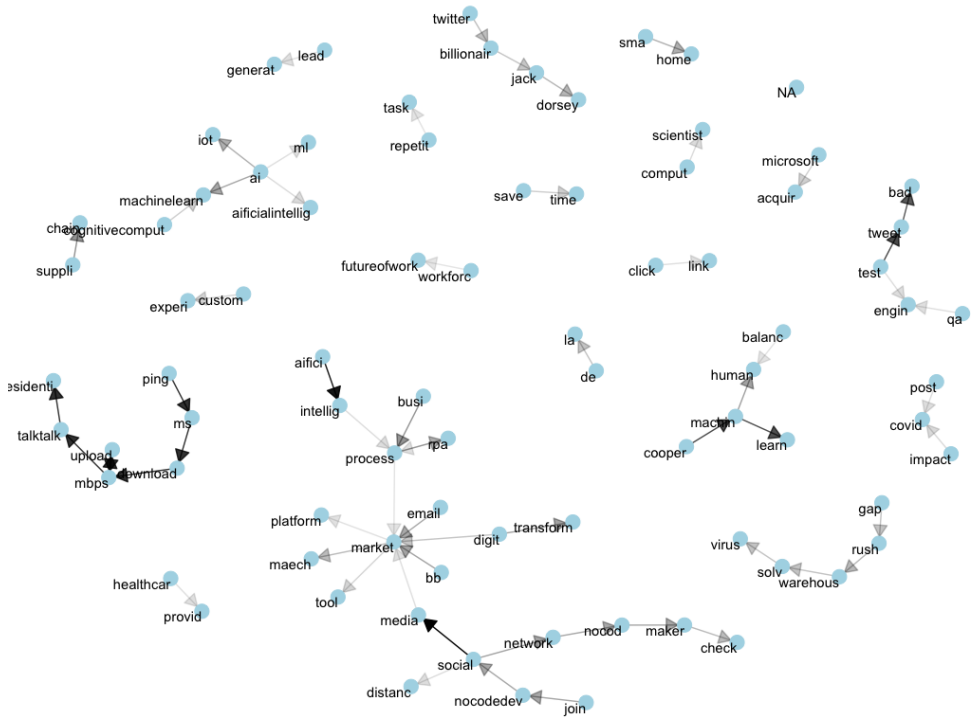

Figure 14: Trigram-network for automation

3.10 Network: Migration

Again we can see that some nodes managed to surpass the filtering process. Migration and labor are terms that appeared at the center of the chart. Compared to the network of automation, one can see a stronger proximity to one tree. Few topics fall out of the realm of this tree such as the node with climate change.

One should point out that here numerous nodes focus on other countries and areas. The EU is mentioned twice and the ukand germany also are centrally located. It appears this dicusssion is broader spread. To look at the narrative within the US, one would need to filter for location which was not done here as that was not the intend. One should note that this is difficult irrespective due to the lack of geographical data entailed in the tweets. Numerous observations entail missing values for this coloumn.

Next, we can look at time trends.

3.11 Time series Plot

This section introduces an example of how to model changes in twitter frequency over time. Prior research has used this type of plot to visualize historical events. Further research has could use these methods to measure polarity-changes over time. This study could be extended to study for a longer period of time to study the impact of the economic lockdown in the long run.

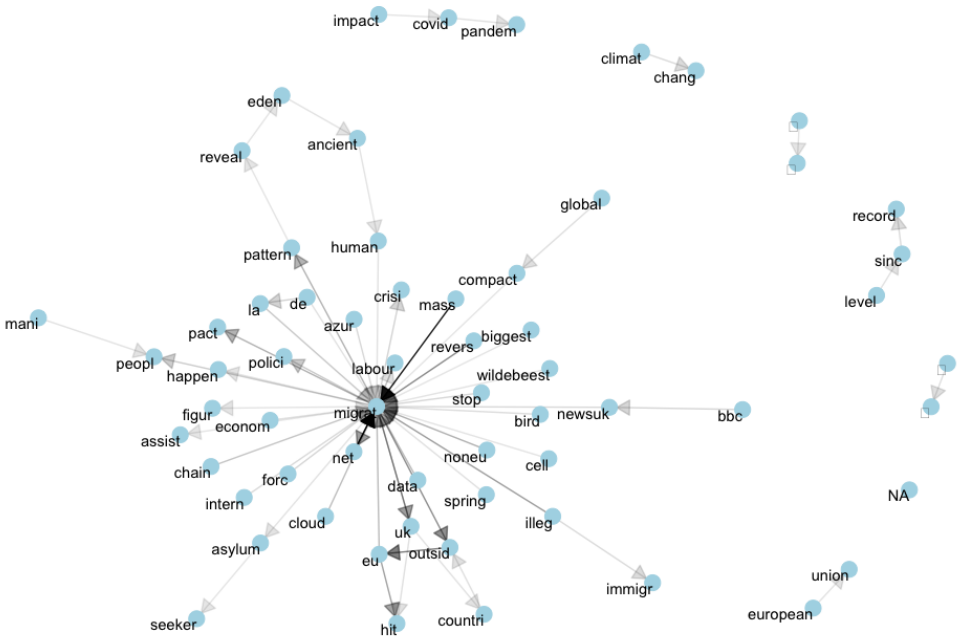

Figure 15: Trigram-network for migration

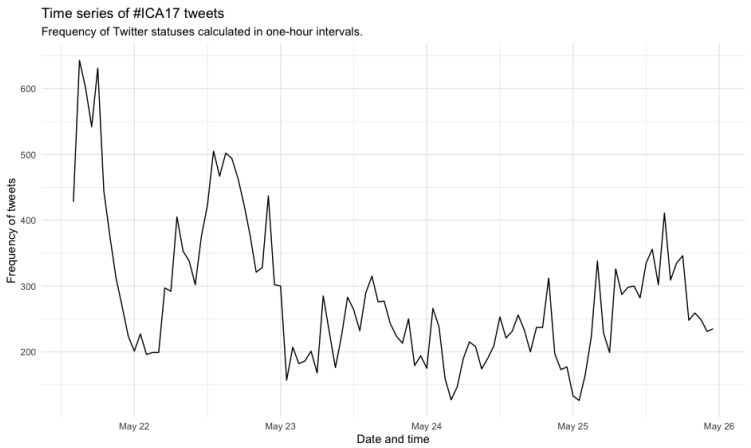

Figure 16: Time series for automation

The second time series model shows how frequency of treets on migration changed. One can see that there is spike between the 21st and 22nd of may. At this period of time the US decided to extend their border policy. This illustration shows that looking at twitter data, one can examine different historic events.

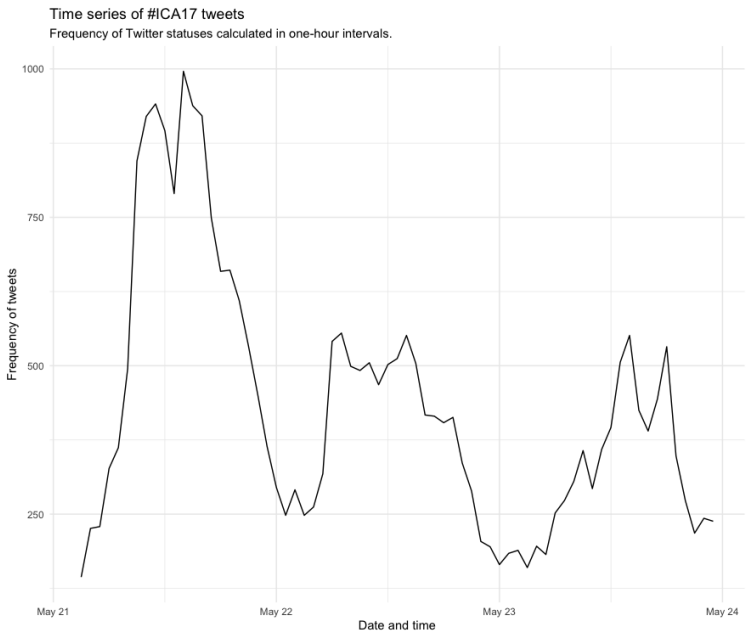

Figure 17: Time series for migration

Furthermore, future research could look at changes in polarity over time. Looking at the

It may make sense to look at the specific topics. Some sentiments could be taken out of context.

4 Conclusion

The assumption that technological change will continue in a linear manner is unlikely. Technological change is volatile as are the repercussions of technology. The assumptions made by Webb (2019) are insightful for future research. Irrespective, next to the concerns expressed above Webb (2019) also imposes some limitations that need to accounted for. The study was unable to embed wage decline, unemployment or movements. Moreover, other factors such as political factors were excluded from the narrative. Ultimately, the descriptive model is interesting, but future research needs to complement these insights with mix methods tools to get insights into the exact nature of these trends. Irrespective, evidence suggest that the displacement effect and productive effect do interplay reshaping the landscape of the labor market. Future research on how to accommodate these displaced workers is pivotal. Especially given the increasing burden to pursue further education, policy makers need to provide guidance for both high, medium and low skilled workers (frank_toward_2019). Increasingly flexibility and technological change are part and parcel of our workforce and intend to stay. Automation, artificial intelligence and the implementation

of further software are inevitable predicaments of the labor market. More research needs to focus on disentangling the relationship between technology and the rate of substitution of work. While this relationship may appear linear on first sight, it clearly is bifurcated. As mentioned by Webb (2019), different technologies have a different impact on various groups. Technological evolution and automation do not see class or color. While AI may cause reduction in cost, demand potentially could also increase, creating higher labor demand. This trend has not been studied sufficiently.

To complement these findings, this study attempted to disentangle fears and threats on social media. Data on common political topics was collected to get insights into how political views might have shaped and what the current center of attention is. Surprisingly, findings hint to a mixed perception of the recent unemployment wave. There were numerous indicators that the stance also entailed hope rather than hate and blame. Further, automation did not manifest as a threat. This research suggests that results can substantially deviate between topics. Irrespective, the dictionary methods provide corroborating evidence that text analysis allows one to obtain a glimpse into social media trends. One should note that the results should be interpreted with caution. Future mixed methods models, text classification regressions, generative models and word embeddings illustrate sizable potential within the literature. These methods need to find further application in social sciences, to allow for more concise evidential support.

Jobs entailing a larger share of interpersonal and cognitive components are less prone to displacement due to technological innovation. Workers within low skilled jobs are most likely to be exposed by robots. Workers within middle income jobs are most likely to be exposed by software. AI has an impact of jobs of the highly skilled. Machine learning is designed to solve complex decision making, predicitive modeling and other advanced challenges. Machine learning is an indication that most sectors of our economy will be exposed to technological change. Policy makers need to accomodate this universal impact. Further, the sentiment analysis allows us to get more insights into the feelings towards work, AI and immigrants. One should note that these results are merely desciptive and dont ensure causality.

5 Discussion

How we think about the future of work and what we think about threats to employment will shape the future political and inevitably, economic landscape. This paper examined threats to the labor market among the general public using twitter data. Examining common fears to employment, sentiments were strongest amongst tweets related to migration and China. Why does this matter? As mentioned by Autor et al. (2016), economic hardship has reshaped voting behaviour, hinting towards further partisanship. What we fear and what we perceive as a threat, shapes the narrative of politics and policy making. In the rising age of political opportunism and political entrepreneurship, it is important that the voting population focuses on actual threats and fears rather than polarizing towards different narratives and following a resource allocation war between different groups. As pointed out, numerous scholars have advocated for an endogenous evolution of technology. Technological evolution is also a predicament inevitable to the future of work. Hence, a critical stance

towards the future of work is needed for a sustainable economic growth trajectory. Therefore, our political and institutional corridor needs to accommodate a sustainable automation pace. When looking at the social media data, excess automation does not manifest as a threat to employment. While at this point it is not surprising that the economic lockdown and news on China manifest in the center of attention, future research needs to look into how technology is perceived in the long run. A healthy relationship to automation and technology is pivotal to a growth sustaining society. Rather than polarization attention to political topics such as foreign policy or migration, real threats needs to be disentangled and communicated. To sustain sustainable technological and ultimately economic growth, academia needs to communicate their ideas allowing them to unfold within political debate and manifest in voting behavior. As pointed out by Autor et al. (2016), descriptive evidence hints towards the fact that trade with china has led to an increased polarization in states which have suffered economic hardship. Future research needs to disentangle correlation and causation when it comes to technology and unemployment. This study did not intend to provide causal analysis, but merely provided descriptive evidence towards what sentiments people have and what people are communicating via twitter. Further, long-run changes towards certain topics need to be examined to further improve communication and understand common fears and threats. Given the rise in relevance of social media more research needs to be done on the usage of bots to spread ideological predicaments.

6 Appendix

6.1 Tables:

Table 7: **Occupations with highest and lowest exposure to robots**

Least Exposed	Most Exposed
Payroll clerks	Forklift drivers
Clergy	Operating engineers of cranes, derricks, etc.
Art/entertainment performers	Elevator installers and repairers
Correspondence and order clerks	Janitors
Eligibility clerks[1]	Locomotive operators: engineers and firemen

Table 8: Occupations with highest and lowest exposure to Software

Least Exposed	Most Exposed
Barbers	Broadcast equipment operators
Podiatrists	Water and sewage treatment plant operators
Subject instructors, college	Parking lot attendants

[1]Refined to eligibility clerks for government programs

Least Exposed	Most Exposed
Art/entertainment performers	Packers and packagers by hand
Mail carriers for postal service	Locomotive operators: engineers and firemen

Table 9: Occupations with highest and lowest exposure to artificial Intelligence

Most Exposed	Least Exposed
Clinical laboratory technicians	Animal caretaker, except farm
Chemical engineers	Food preparatiion workers
Optometrisits	Mail carriers for postal service
Power plant operators	Subject instructors, college
Dispatcher	Art/entertainment performers

Table: **Regression table summary of exposure on wages (1;3) and employment (2;4)**

	(1)	(2)	(3)	(4)
Exposure	-0.22***	-0.16***	-0.04***	-0.14***
	(0.03)	(0.03)	(0.01)	(0.02)
Offshorability	-2.29***	2.02***	-0.87***	2.66***
	(0.50)	(0.55)	(0.28)	(0.53)
Medium education	9.52***	-1.20	11.80***	6.19***
	(1.67)	(1.54)	(0.93)	(1.35)
High Education	27.73***	14.42***	32.75***	22.26***
	(2.01)	(2.40)	(1.24)	(1.91)
	-0.07***	0.04***	-0.07***	0.03***
Wage	(0.01)	(0.00)	(0.00)	(0.00)
	0.00***	-0.00***	0.00***	0.00***
Wage squared	(0.00)	(0.00)	(0.00)	(0.00)
R^2 – Adjusted	0.163	0.147	0.168	0.210
Industry FEs	Yes	Yes	Yes	Yes
Observations	6,708	14,065	18,975	36,070

Figure 18: Regression replication table

6.2 Topic Modelling

6.2.1 Network - China

The network tri-grams illustrate multivariate topics, relating to china in public debate. Here, due to the inability due to shortcomings in the cleansing process, symbols that should not be in the data, remained present. These may be chinese symbols that didnt filter out. Irrespective the trigram is presented here for reference.

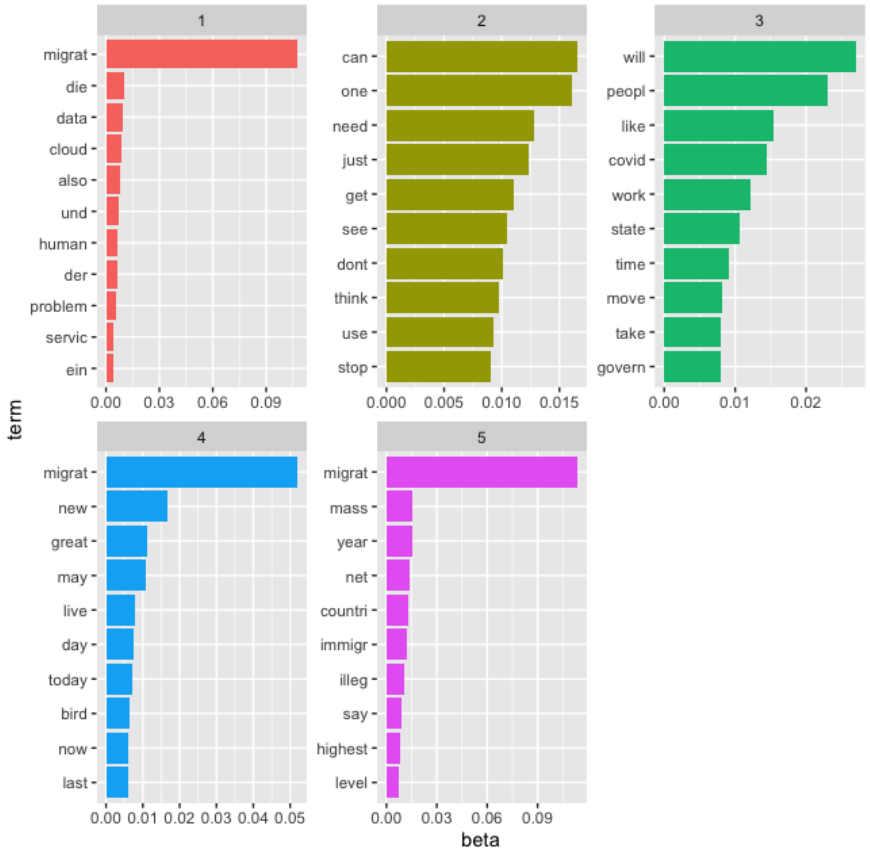

Figure 19: Topic modelling using gibbs sampling for migration

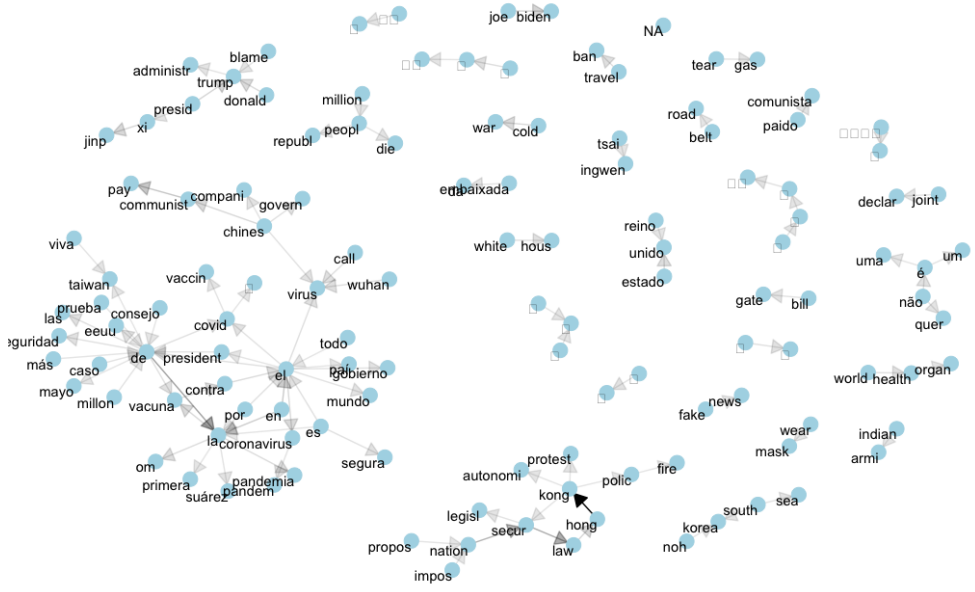

Figure 20: Networks for China

Some of these topics are the separation movement in Hong Kong and Taiwan. Further, "indian" and "armi" hint towards the border conflict in Ladakh.

6.2.2 Networks: Employment

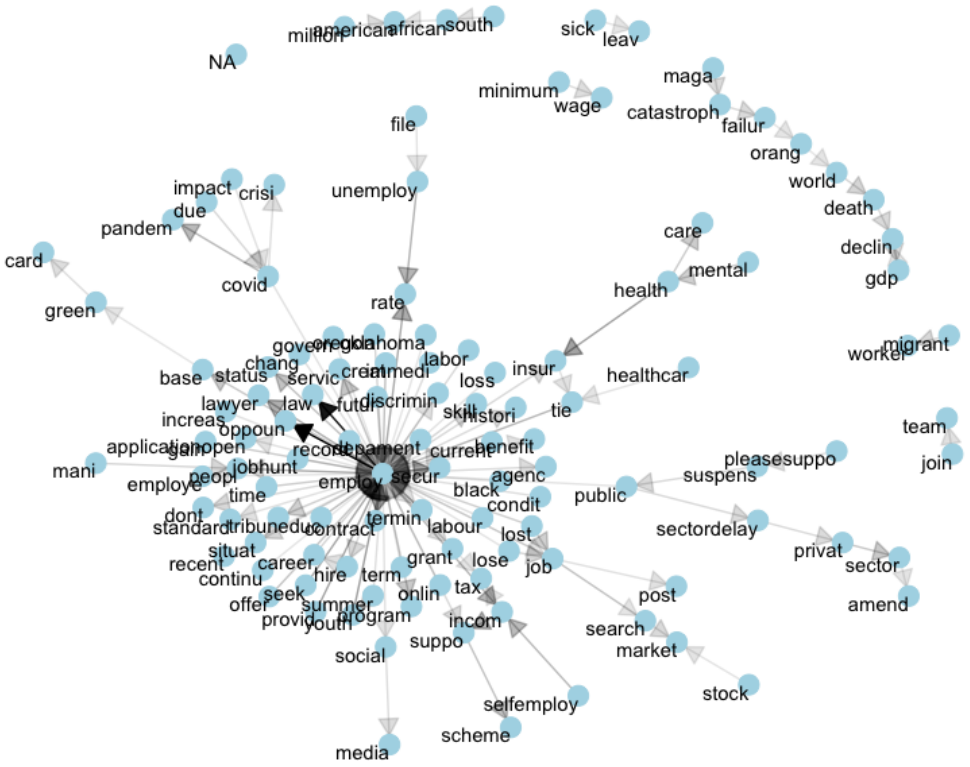

Figure 21: Networks for Employment

6.2.3 Mapping tweets

This section provides a map of the location of the twitter useres tweeting content related to the keyword migration:

As we have not specified tweets to be solely from america, but did use english tweets, we can see in what other countries this topic has gained prominence on twitter.

One can see that a considerable amount of tweets are coming from the UK and Africa. On the other hand, one can see that tweets about automation are centered differently compared to migration. One can see that most of the tweets orginate from the US. Irrespective some outliers do stick out. One can see that there are a sizable number of tweets from india. The employment chart corroborates the findings suggested by the LDA.

References

Acemoglu, Daron, and Pascual Restrepo. 2018a. "Artificial Intelligence, Automation and

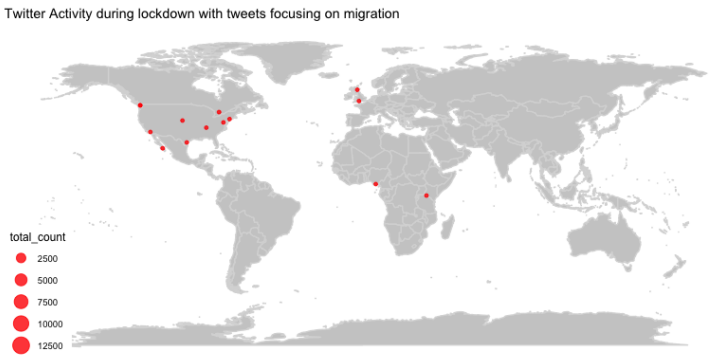

Figure 22: Location of tweets for Migration

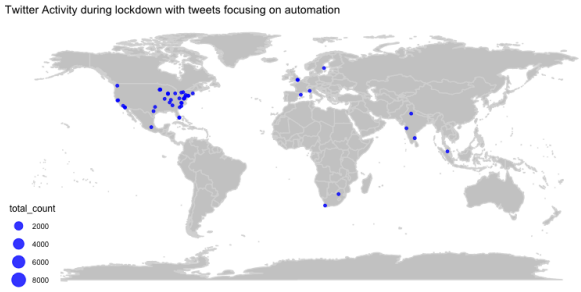

Figure 23: Location of tweets for Automation

Work." National Bureau of Economic Research.

———. 2018b. "Modeling Automation." In *AEA Papers and Proceedings*, 108:48–53.

———. 2019. "Automation and New Tasks: How Technology Displaces and Reinstates Labor." *Journal of Economic Perspectives* 33 (2): 3–30.

———. 2017. "Low-Skill and High-Skill Automation." Working Paper 24119. National Bureau of Economic Research. https://doi.org/10.3386/w24119.

———. n.d. "Robots and Jobs: Evidence from US Labor Markets." 90.

Acemoglu, Daron, and James A. Robinson. 2015. "The Rise and Decline of General Laws of Capitalism." *Journal of Economic Perspectives* 29 (1): 3–28.

Athey, Susan. 2018. "The Impact of Machine Learning on Economics." In *The Economics of Artificial Intelligence: An Agenda*, 507–47. University of Chicago Press.

Autor, David, David Dorn, Gordon Hanson, and Kaveh Majlesi. 2016. "Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure." w22637. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w22637.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics* 118 (4): 1279–1333.

Bessen, James E. 2017. "Automation and Jobs: When Technology Boosts Employment." *Boston Univ. School of Law, Law and Economics Research Paper*, no. 17.

Blei, David M. 2003. "Latent Dirichlet Allocation," 30.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. 2017. "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association* 112 (518): 859–77. https://doi.org/10.1080/01621459.2017.1285773.

Ferri, Fernando, Alessia D'Andrea, and Patrizia Grifoni. 2017. "An Integrated Methodology for Approaching Sentiment Analysis in Business Domain." *International Business Research* 10 (9): p1. https://doi.org/10.5539/ibr.v10n9p1.

Frank, Morgan R., David Autor, James E. Bessen, Erik Brynjolfsson, Manuel Cebrian, David J. Deming, Maryann Feldman, Matthew Groh, José Lobo, and Esteban Moro. 2019. "Toward Understanding the Impact of Artificial Intelligence on Labor." *Proceedings of the National Academy of Sciences* 116 (14): 6531–9.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–74.

Gregory, Terry, Anna Salomons, and Ulrich Zierahn. 2016. "Racing with or Against the Machine? Evidence from Europe." *Evidence from Europe (July 15, 2016). ZEW-Centre for European Economic Research Discussion Paper*, no. 16.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3):

267–97.

Grün, Bettina, and Kurt Hornik. 2011. "**Topicmodels** : An *R* Package for Fitting Topic Models." *Journal of Statistical Software* 40 (13). https://doi.org/10.18637/jss.v040.i13.

"Is Automation Labor-Displacing? Productivity Growth, Employment, and the Labor Share." n.d. Accessed May 29, 2020. https://www.nber.org/papers/w24871.

Melluso, Nicola, Silvia Fareri, Gualtiero Fantoni, Andrea Bonaccorsi, Filippo Chiarello, Elena Coli, Vito Giordano, Pietro Manfredi, and Shahin Manafi. 2020. "Lights and Shadows of COVID-19, Technology and Industry 4.0." *arXiv Preprint arXiv:2004.13457*.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Niewiadomski, Robert, and Dennis Anderson. 2020. "The Rise of Artificial Intelligence: Its Impact on Labor Market and Beyond." In *Natural Language Processing: Concepts, Methodologies, Tools, and Applications*, 1298–1313. IGI Global.

Piketty, Thomas. n.d. "About Capital in the 21st Century," 15.

Salomons, Anna. 2018. "Is Automation Labor-Displacing? Productivity Growth, Employment, and the Labor Share." National Bureau of Economic Research.

Sendhil Mullainathan; Jann Spiess. n.d. "Machine Learning: An Applied Econometric Approach - American Economic Association." Accessed May 29, 2020. https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87.

"Sentiment in Twitter Events - Thelwall - 2011 - Journal of the American Society for Information Science and Technology - Wiley Online Library." n.d. Accessed May 28, 2020. https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21462.

Silge, Julia, and David Robinson. 2017. *Text Mining with R: A Tidy Approach*. "O'Reilly Media, Inc.".

"U.S. Jobless Claims Pass 40 Million: Live Business Updates." 2020. *The New York Times*, May. https://www.nytimes.com/2020/05/28/business/unemployment-stock-market-coronavirus.html.

Webb, Michael. 2019. "The Impact of Artificial Intelligence on the Labor Market." *Available at SSRN 3482150*.