Project 6: Social Media Analysis
CS-C4100 - Digital Health and Human Behavior

Anonymous Author

1 Introduction

This report will present an in depth look at the *Tweets about Covid-19 all over the world* dataset (Khetlani, 2021). We will go over cleaning methods for social media content (Han and Baldwin, 2011)(Section 3.1) and cover the distributions in the dataset (Section 3). Perform machine translation (Jehl, 2010) on selected languages (Leite et al., 2020) (Section 4.1) and then we will perform an sentiment analysis. We will connect the sentiment to the time of posting to evaluate if there is a different sentiment depended on the time of the day. The Analysis results in clear activity pattern but no observable dependence of time and sentiment. We also find that there are slightly more negative tweets on average across all languages.

A comparison between topics extracted for different languages using TF-IDF (Ramos et al., 2003) is performed (Section 4.3). To visualizes what speakers of different languages talk about we will use word clouds. We find that people speaking different languages use similar words in the dataset and we can not find a vase differences between the languages when only looking at the frequency of used words. The report is motivated by the fact that most people are not able to understand more than two languages and facing a problem of global scale, like the covid-19 pandemic, it is important to get an insight into what people speaking languages we do not understand talk about.

We also take a look at the difference in twitter usage (Section 4.4) of different languages. We will compare the dataset to a study of 2011 (Weerkamp et al., 2011) to see if the way Twitter is used changed over the years or when looking at a topic centered dataset instead of general twitter traffic. The result is that the majority of languages now use more Hashtags but less Mentions. This could be due to the topic centered dataset. The only language where is trend is inverted is German.

2 Problem Formulation

This report will discuss the structure of the underlying, multi-language Dataset. We will present observations within the contextual differences of Tweets published in different languages and tackle the main question if there are significant differences between different languages. The differences of interest to this report are the variation in publishing times and sentiment. We want to understand if there are significant variations in these aspects and discuss the findings with the background that the tweets revolve around a pandemic.

Using topic recognition we would like to understand what exactly is talked about in the dataset. Further splitting tweets into language groups we want to observe if there are noticeable differences between the topics of languages. To understand how people in different

countries, speaking different languages, deal with the same problem here the Covid-19 pandemic it is important to not only regard tweets of one (English) language since most people will express themselves in their mother tongue. With this report I want to give an insight into what people tweeting in non-English languages care about and find the similarities and differences in that regard.

It was previously shown that speakers of different languages use Twitter mentions and hashtags at very different frequencies (Weerkamp et al., 2011). We will calculate the statistics for the dataset and see if observations obtained in 2011 still hold. The motivation for this analysis lies within the vastly different spread of Covid-19 around the world and therefor a difference in communication could be expected based on the nationality of a users. We want to see if these assumptions can be verified and if not if other patterns emerge.

3 Dataset Description

The used Dataset called *Tweets about Covid-19 all over the world* (Khetlani, 2021) was constructed by Komal Khetlani. The Dataset was published through kaggle.com. It is constructed using Twitter data where each data point resolves around a so called *tweet*. Twitter uses the pattern "@TwitterUser" to represents that this message is a reply to the user (a mention) called "TwitterUser" and tags like "#aalto" provided by the user for this message are so-called hashtags. Tweets are limited to 280 characters. The dataset contains around 804 thousand data points in its initial form. Table 1 shows the initial columns of the Dataset along with the present data type and a short description of the given column. It is supposed to give a first overview over the raw dataset.

Column name	Data type	Description
id	Integer	Unique identifier of the Tweet
created_at	String	Date of creation of tweet
date	String	Date of the Tweet
time	Integer	Time of the Tweet
timezone	Integer	Timezone of the Tweet
place	Dictionary	location coordinates
tweet	String	The tweet
language	String	Language of the tweet
replies_count	Integer	The number of replies to that Tweet
retweets_count	Integer	The number of retweets
likes_count	Integer	The number of likes to that Tweet
hashtags	Array	The hashtags used
cashtags	Array	The cashtags used
retweet	Boolean	Has any retweet done on the tweet
video	String	Video url
thumbnail	String	Link to Thumbnail image

Table 1: Initial Dataset columns

Firstly lets discuss the columns which were not modified. The column *id* is a unique identifier for each tweet and will serve as an index. The column *created_at* carries the time of the

tweets publishment with an accuracy of seconds. The only modification to the column is the conversion to pandas datetime format which changes only the usability and not the present information. This column will be used for everything revolving around time. Information about the tweets language is stored in *language*. Here the ISO 639-1 standard for language codes is used. Since we are interested in differences between languages we needed this column to group tweets later. The actual tweets are stored in the column *tweet*. While we will perform some cleaning of the tweets (Section 3.1) the original tweets will not be modified and rather additional columns carrying the changes are created. To later perform an analysis based on hashtag usage (Section 4.4) the column *hashtags* will stay. Since the original dataset is a csv file the arrays encoded in the strings of the hashtag column will be loaded using the json module of python.

At this point we discussed the columns *id, created_at, tweet, hashtags* and *language*. All other columns will be disregarded with the following reasoning. The columns *replies_count, likes_count* and *retweets_count* will not be used in any part of the exploration and can therefore be dropped. The research problem does not benefit from the metricise. The columns *date* and *time* carry redundant information to *created_at*. Both combined present the same information as just the one column which was picked to stay. Further the *timezone* column is 0 for all entries and therefore carries no information at all. The same applies to the *retweet* column with the slight exception that one entry is null and all others are false. Again this column carries no information and will be dropped. While *place* has entries for roughly 0.1% of all entries it is not connected to the research question nor could it be used with such a small number of non null values. An analog argument is made for the *cashtags*. Again less than 1% of the entries carry values. Finally the columns *video* and *thumbnail* are again not interesting towards the research question.

To conduct an analysis on hashtags and mentions later (Section 4.4 we will create the column *mentions* in which an array of all usernames that the corresponding tweet has mentioned is stored. The column *translated_tweets* is created during the translation process (Section 4.1).

3.1 Text cleaning

In this section we will look over the cleaning steps that where done to create the added column *cleaned_text*. Social Media content from twitter is known to be noisy (Han and Baldwin, 2011). Through cleaning we want to reduce the noise and increase the accuracy of our resulting analysis. This is archived by removing hyperlinks, user mentions, hashtags, numbers as well as line breaks and other white space characters. These steps can be executed before the translation since they are independent of the underlying language. The final step of the per processing is the removal of stop words. Stopwords are words that are commonly used in a language but have little or no meaning on their own, such as articles, conjunctions, and prepositions. These words are often filtered out of text when performing natural language processing tasks, as they can interfere with the meaning of the text and make it more difficult to analyze. This step is done after translating the tweets. In the implementations the stopword corpus of nltk was used and only words not present in that corpus are kept. As a last step we check for tweets with less than five words and remove all that carry too little information. The already mentioned columns as well as the three new columns conduct the dataset after per processing.

3.2 Language selection

In the original dataset there are 64 unique languages present. Notice that the dataset does not provide information how language detection was performed. We will select four for the upcoming analysis. The selected languages are English, Spanish, German and French due to there relatively high number of tweets and the availability of research. Figure 1 displays the distribution of all present languages. Most languages only contribute a small percentage to the dataset with 61 languages contributing less than 5%, 55 languages contributing even less than 1% and 35 languages contributing less than 0.1% of the total number of tweets. To still visualize languages with a small number of tweets a logarithmic scale was used on the y-Axis. The plot was created using seaborn.

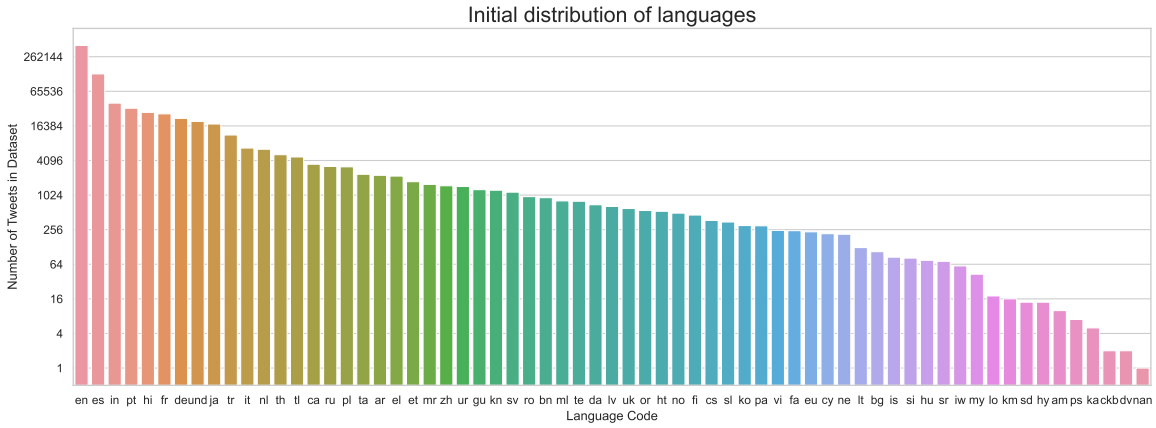

Figure 1: Frequency of languages in the dataset.

3.3 Dataset statistics after preprocessing

After all the steps descried in sections 3.1 and 3.2 we are left with a dataset containing around 509k tweets and four distinct languages. Table 2 shows the finished structure with explanations for each columns. The described dataset can be found as a pickle file in the submission files. The languages are still identifyed using the ISO 639-1 standard and the contained values are en: English (411.738 Tweets), es: Spanish (131.380), de: German (22.034 Tweets) and fr: French (26.506 Tweets).

Column name	Data type	Description
id	Integer	Unique identifier of the Tweet
created_at	String	Date of creation of tweet
tweet	String	The tweet
cleaned_tweet	String	The tweet with out URLs and other symbols
translated_tweet	String	The tweet in English
language	String	Language of the tweet
hashtags	Array	The hashtags used
mentions	Array	The usernames that are mentioned

Table 2: Dataset structure used for analysis

The dataset contains tweets in a time window of four days ranging from the 20st of April (afternoon) to the 24th of April 2021 (evening). All times discussed from here on are with reference to the UTC timezone.

4 Methods

By now we understand the structure of the Dataset and discussed the executed cleaning methods. This section will cover the used methods starting with translating the non-english tweets (Section 4.1) further we will perform sentiment analysis along with visualizing the polarity and times of tweets grouped by languages (4.2). We then perform topic detection and visualize it using word clouds (Section 4.3). Lastly we will examine if the findings of (Weerkamp et al., 2011) hold for a dataset created about 10 years after their study (Section 4.4).

4.1 Machine Translation

The Dataset consists of over 60 languages (Section 3.3). The portion of English Tweets is roughly 50%. Since my research problem revolves around the differences in languages I need to be able to analyse not only English but also Spanish, French and German Tweets. Most research is conducted in English and while there exist methods and models for the other mentioned languages a streamlined process would be more convenient to analyse since the output format is than equal for every language.

The translation itself can be conducted using a already existing translator like Google Translate DeepL or LibreTranslate. It has been shown that for topic detection or other bag-of-words observations translated tweet are a viable option (Leite et al., 2020). Google Translate and DeepL both have subscription based APIs and while DeepL is free for less than 500.000 characters per month the non English tweets that need to be translated total around 40 million characters. The chosen translator is LibreTranslate due to the ability to host it yourself and therefor not being limited in the number of characters. To perform translations LibreTranslate launches a local server that can be accessed via http requests. The server already has the option to use multiple threads which provides parallelization for the front end. We split the dataset into four equally big pieces and use multiple notebooks to speed up the translation process. Using said notebooks in parallel the whole translation process took around 28 hours to compute. While this is a considerable amount of time, this approach could be scaled to use more threads archiving managable speeds even for very big datasets.

4.2 Sentiment Analysis

TextBolb is a open source python library which provides a variety of language processing tools. We will use there sentiment analysis module to analyse if the sentiment in tweets is dependent on the original language or the time of creation. TextBolb uses a rule based approach to compute the polarity of a given sentence. To obtain the sentiment a average over all sentences in a give tweet is taken. If the average is greater than 0 a positive sentiment is conducted vise versa a negative label is set. Other articles have shown that there are time dependent events on twitter, which influence sentiment (Thelwall, 2014). In Figure 2 the creation times and sentiments in Tweets grouped by languages are displayed. The plot is a modified violineplot from the python package seaborn. The top part of each violine represents the amount of negative tweets at that point in time analog to this the lower part represents the amount of tweets tagged with a positive sentiment.

Column name	Data type	Description
sentiment	String	positive/negative
polarity	Float	Polarity ratio
timestamp	Integer	Unix timestamp

Table 3: Added columns to dataset for sentiment analysis

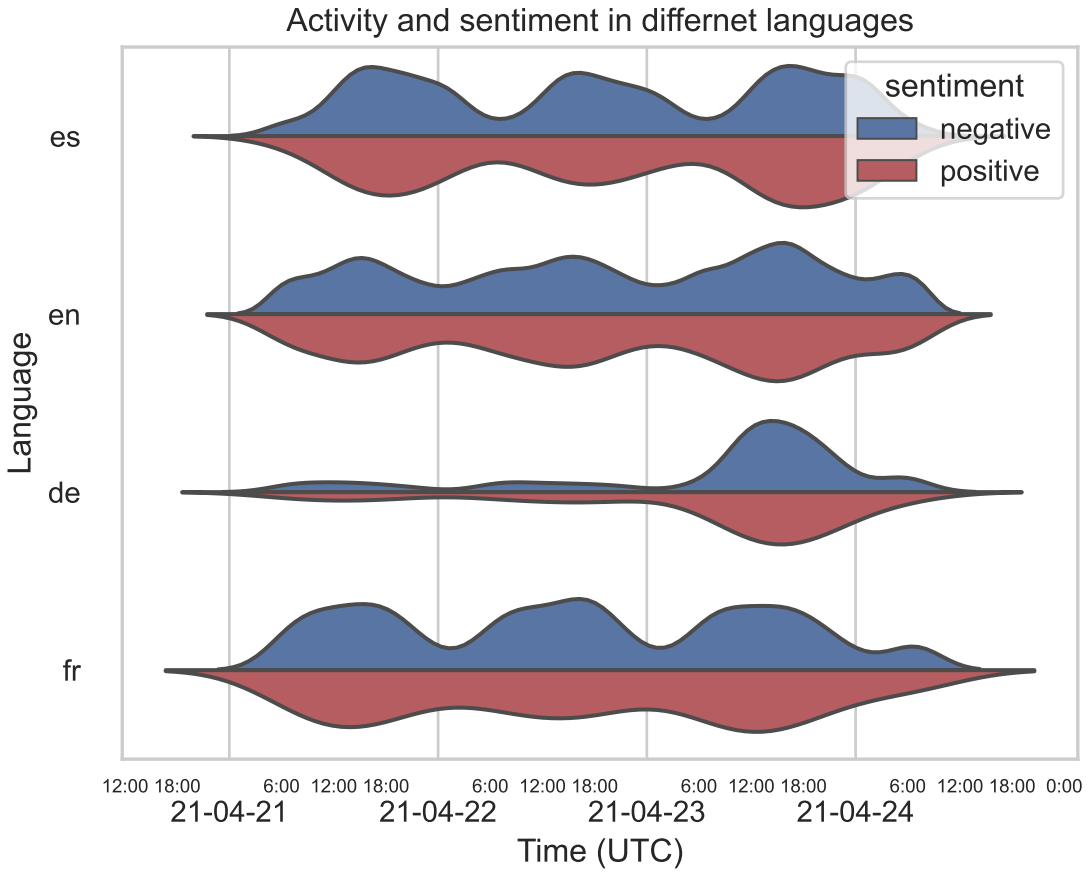

Figure 2: Creation Times and sentiment in Tweets grouped by languages.

4.3 Topic Detection

TF-IDF stands for Term Frequency Inverse Document Frequency. Its a messure to conduct which words are the most relevant for some document by looking at their frequencies (TF) as well as their uniqueness to the given document within all documents (IDF). To calculate the scores the following formulas are used:

$$tf = \frac{\#term\ in\ document}{\#all\ terms\ in\ document},\ idf = \log\left(\frac{\#documents}{\#documents\ term\ is\ in}\right),\ tfidf = tf \cdot idf \quad (1)$$

In the implementation scikit-learn is used to calculate the tf-idf matrix. A document is constructed as the aggregation of all tweets of one language. This means we iterate over all tweets of a given language and concatenate them. For improved quality we also lemmatize each word. Lemmatization is the process of *"reducing inflectional forms and sometimes derivationally related forms of a word to a common base form"* (Manning et al., 2008). This

should increase the quality since we have less words that are just derivations of each other. To extract the most relevant words for each document we pick the 25 words with the highest tf-idf score. To select these top 25 values the python module heapq is used. Rather than iterating over all values 25 times to extract the biggest values heapq uses a heap to perform the same operation in a faster time. The number 25 is chosen to ensure the it is possible to see differences while not overloading the visualization with too many words. To visualize the words the python module WordCloud is used. The module has the option to generate a wordcloud from given frequencies and is therefore conveniently applicable to the already extracted data. The resulting plot is figure 3. To group all wordcluds in one plot we use matplotlibs subplots on a two by two grid.

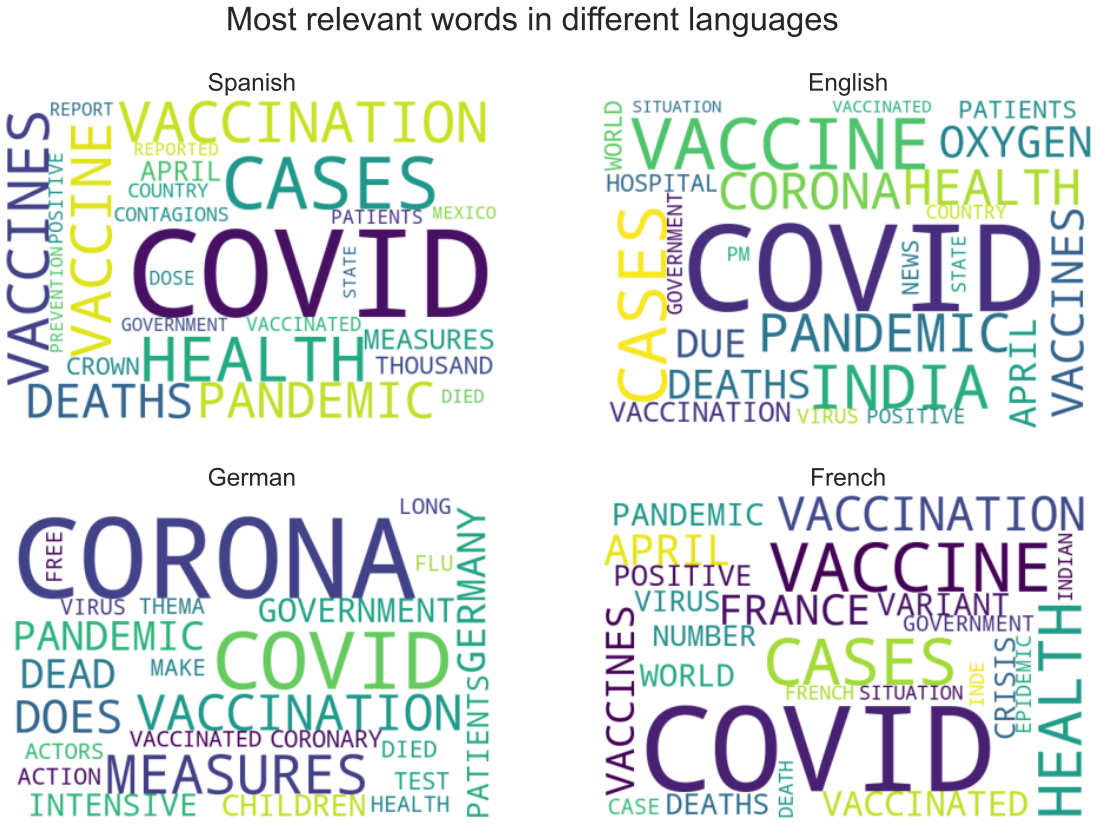

Figure 3: WordClouds of different languages. Created using TF-IDF and translated tweets grouped by language

4.4 Hashtag usage

In past work from 2011 (Weerkamp et al., 2011) the authors observed noticeable differences in the use of hashtags and mentions from speakers of different languages. During pre-processing (Section 3.1) the column *mentions* has been created. Combining it with the column *hashtags* present in the original dataset we can extract the same metrics as the mentioned article. The authors chose eight different languages which are Dutch (nl), English (en), German (de), French (fr), Indonesian (in), Japanese (ja), Portages (pt) and Spanish (es). The ordering is kept in line with the comparison study. Figure 4 shows the results of the dataset in direct comparison to the results of Weerkamp et al. The barplots in the first row give a side by side comparison for each language regarding the ratio of used hashtags and mentions. The scatter plot in the second row shows where the different languages have moved compared the observations obtained by Weerkamp et al. All plots are created using matplotlib.

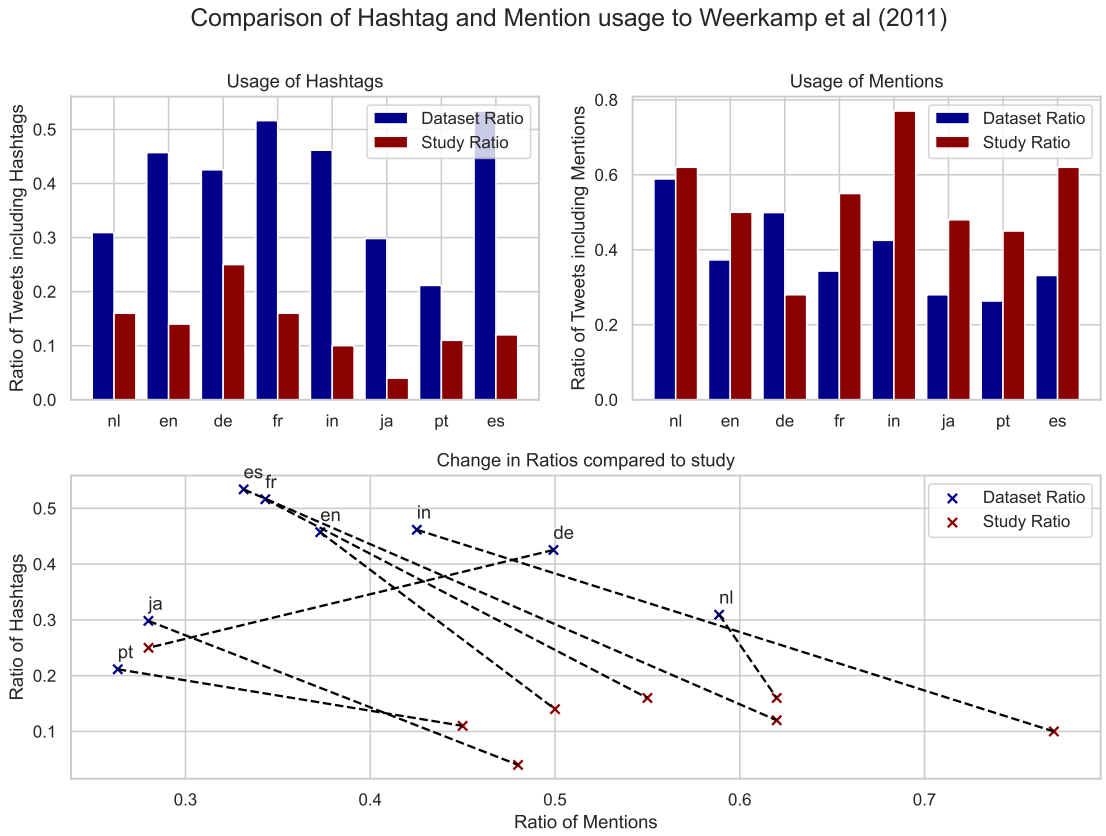

Figure 4: Hashtags and Mentions in different languages in comparison to known stats from 2011.

5 Results

In this section we will go over the results and figures generated in the previous section. We will go over the findings in detail and discuss implications as well as limitations.

5.1 Implications

The sentiment analysis performed in section 4.2 and the resulting plot displayed in figure 2 yield the following observations. Firstly we notice the increased amount of German posts on the 21st of april compared to the other days where only very few German post are present. This spike in activity is probably due to the data creation process rather than a real world phenomenon though I can neither prove or disprove this since the infomation on how exactly scraping was performed is missing in the dataset description. The activity curves of all languages behave similarly. We can clearly see a drop in activity during the nightly hours and for all languages the activity peaks in the afternoon expect spain where the peak is at noon. Conducting insights on the sentiment is not as clear as the active hours. Clearly the shapes are similar in shape but there are is a higher number of post with negative sentiment in all languages. The biggest difference of positive and negative sentiment has been observed in German where the smallest differences is present in English. In general the active times and sentiments do not seem to be dependent of the tweets language.

We will continue with the WordClouds created in section 4.3. The main observation is again that the different languages are fairly similar in the topics they discuss. We are able

to catch sings of the original language by the countries mentioned. The Spanish cloud includes 'Mexico', the French cloud 'France' and the German one includes 'Germany'. The most noticeable difference is probably the term 'Corona' in the German cloud. This can be explained by the fact that Germans and German media mostly refers to the Covid-19 pandemic using the term 'Corona'. We can also observe that all clouds include terms about vaccination. Considering the time of creation (21-24 April 2021) is about the time when vaccines became available to most people in the world it makes sense that this is a heavily talked about topic. In short we can conclude that the topic discussed most in the dataset are vaccines and there is no appeared difference between the languages.

Connecting the observations from the sentiment a possible interpretation is that the availability of vaccines provide hope (positive sentiment) but at the same time covid-19 causes deaths everyday (negative sentiment). In more general terms we can conclude that it does not make a significant difference which of the observed language someone speaks when looking at the broad view provided in Figure 2.

In section 4.4 we compared the usage of hashtags and mentions of the tweets in the dataset grouped by languages to a study from 2011 (Weerkamp et al., 2011). It is clear that the usage of mentions and hashtags has changed. In all languages hashtags have seen an increased used and further all languages but German use less mentions compared to the findings of the study. In the scatter plot we can see the the 'top-right' moving trend of languages. We could interpret this behaviour by the fact that the dataset is based around a topic so people talking about the given topic are more likely to use related hashtags to there tweet than mention someone they know or a celebrity.

5.2 Limitations

Certainly there are limitations regarding the presented analysis. Firstly most of the results rely on the translation quality. While we discussed that for bag of words style analysis translation should not be a problem it would be naive to think that the translations are perfect. Especially since Twitter data is as noisy as it is. Further the dataset is based around a single topics: Covid-19. This means it makes sense that speakers of different languages talk about similar situations since Covid is a pandemic. To get more insights into relevant topics a long term, topic independent dataset could yield more insights.

The comparison work for the hashtag/mention analysis has a relatively small size of 1000 tweets per language. Due the small size they can not guarantee good stability of there findings. Further the dataset used in this report is topic specific and therefore a more similar behaviour regarding mentions and hashtags could be expected. This means the comparison could yield the conflicting information it does not because twitter users actually changed their behaviour but rather the circumstances observed are too different.

Recent research has only proven that rule based sentiment analysis like TextBolb or LIWC (Pennebaker et al., 2001) can not compete against modern deep learning approaches (Fiok et al., 2021). This means the quality of findings is also limited by the methods used for analysis and more powerful models could present findings the are more precise.

5.3 Future Steps

To better understand the differences in languages a embedding based clustering could be used (Sun, 2021). Other work presented Bert (Devlin et al., 2018) based text embeddings that are fine tuned on covid tweets and capable of dealing with multilingual datasets (Müller et al., 2020). Using these modern technologies I would expect to find better differences. Further a dataset that covers a bigger time frame could increase the accuracy.

The underlying dataset is in general not documented well. Neither is explained how tweets where selected (based on keywords or some other method) nor was explained how language classification was done. This is especially problematic since language detection in tweets or noisy short social media text in general it not an easy task. For future analysis a better documented dataset could yield more insightful results. Especially since the dataset inspiration is *"To identify trends in the Tweet"* I would be highly interested in future studies covering more aspects aroud covid. A possibility would be to use sentiment analysis on a dataset starting in 2020 up until 2023 to see how the sentiment regarding covid changed. Even a multi-language analysis would be feasible due to the translation speeds discussed in section 4.1.

Another step could be to include more than the four languages picked by me. TF-IDF might yield clearer results when used on eight languages since the inverse document frequency could better distinguish words which are present in every language. A report that continues the topic could als try to adjust the weights of the TF and IDF part to extract better fitting results with the background knowledge we have about the data.

6 Conclusion & Discussion

The report has show that there seems to be no vast differences between languages regarding the choice of words. Further the patterns found in the sentiment and publishment of tweet show stable pattern across languages. In my personal opinion this is a good sign since difference with regard to covid-19 conversation could implies misinformation. Assume some language does not cover vaccines at all in there communication. This could mean that people are not interested in the possibility of protecting themselves are slowing down a pandemic. If some language would have significantly more tweets with positive or negative sentiment it could imply that either covid-19 is not taken serious or that information is not discuss in a calm way.

The results of the hashtag and mentions analysis also give reasonable results. I find it very interesting to see the shift in the amount of hashtag and mentions rations. In general I think we should be more interested in the behavior of other languages that we can not speak. Not only can we learn social patterns but also see that people from different countries are apparently not that different after all. This observation might help to prevent racism or other forms of hate speech aimed at minorities.

Acknowledgements

I want to tank Talayeh Aledavood for the course Digital Health and Human behaviour. Further I want to say thank you to my fellow students which I got to know during the course.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Krzysztof Fiok, Waldemar Karwowski, Edgar Gutierrez, and Maciej Wilamowski. 2021. Analysis of sentiment in tweets addressed to a single domain-specific twitter account: Comparison of model performance and explainability of predictions. *Expert Systems with Applications*, 186:115771.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 368–378.

Laura Elisabeth Jehl. 2010. Machine translation for twitter.

Komal Khetlani. 2021. Tweets about covid-19 all over the world. https://www.kaggle.com/datasets/komalkhetlani/tweets-about-covid19-all-over-the-world.

João Augusto Leite, Diego F. Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *CoRR*, abs/2010.04543.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. „introduction to informationretrieval"-chapter 2.2.4 stemming and lemmatization. *IR–book/html/htmledition/irbook.html*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. New Jersey, USA.

Daniel X Sun. 2021. *Clustering Tweets via Tweet Embeddings*. Ph.D. thesis, Massachusetts Institute of Technology.

Mike Thelwall. 2014. Sentiment analysis and time series with twitter. *Twitter and society*, pages 83–95.

Wouter Weerkamp, Simon Carter, and Manos Tsagkias. 2011. How people use twitter in different languages.