Virtual Tribes: Analyzing Attitudes towards the LGBT Movement by Applying Machine Learning on Twitter Data

Moritz Bittner[a1], David Dettmar[a2], Diego Morejon Jaramillo[b3], Maximilian Johannes Valta[b4]

[a] Universität zu Köln, Albertus-Magnus-Platz, 50923 Köln

[b] Otto-Friedrich-Universität Bamberg, Kapuzinerstraße 16, 96047 Bamberg

Abstract In this paper we investigate the application of machine learning techniques that allow conclusions from users' behavior and language in Twitter about their attitudes towards the LGBT movement. By using an adjusted procedure of the CRISP-DM process (Cross Industry Standard Process for Data Mining) we create a prediction model and formulate step-by-step instructions for its deployment. We provide the reader with a theoretical background for our research domain and precisely describe the methods that we use. Results show that there are two groups of contrary attitudes towards LGBT community and that the language and behavior of users in the groups respectively differ from each other. Also, we identify word analyses as a valuable mean for prediction. We also apply our model on another dataset to investigate its interspersion with the previous identified groups and demonstrate its effectiveness for predicting attitudes of a single actor in Twitter. In the end we critically discuss our findings and recommend further fields of investigation that are related to our research.

---

[1]Corresponding author: Moritz Bittner, email: bittnerm@smail.uni-koeln.de

[2]Corresponding author: David Dettmar, email: ddettmar@smail.uni-koeln.de

[3]Corresponding author: Diego Morejon Jaramillo, email: diego-sebastian.jaramillo@stud.uni-bamberg.de

[4]Corresponding author: Maximilian Johannes Valta, email: maximilian-johannes.valta@stud.uni-bamberg.de

1 Introduction

Since ancient times tribes have been a popular concept in societies (Pritchard, 2000). Tribes are groups of people that share the same language and values like culture and history. In particular, tribe members exalt their tribe above other tribes and groups, which leads to tribal consciousness and tribal loyalty (Cambridge Dictionary; Merriam-Webster Dictionary). The ancient tribes often lived among each other detached from others. When two different tribes met each other, conflicts were likely to arise and differences in social living, technological developments or values came to light (Apelt, 2010).

Today in times of global convergence these strong differences between tribes' realities and belief systems seem to disappear at first sight. However, due to social fragmentation, diversification, and the development of new communication channels in the field of information and communications technology (ICT), communities that form are not easily detectable. In the following, these communities are referred to as virtual tribes. Like ancient tribes, such virtual tribes define their own truths and live within their tribes' reality (Oliveira and Gloor, 2018). By using different tools, it is possible to identify and collect tribe members for any tribal macro-category which is the goal for an investigation by an analyst (Gloor et al., 2018). Later on, the likelihood of a certain social platform user being member to one of these tribes can be measured by using machine learning techniques.

Holding more than 320 million active users (statista.com, 2016) and 500 million tweets per day (Twitter, Inc.), Twitter is a great source of data that can be used for research. In the past, there have been lots of scientific investigations based on its plurality of accessible data, like e.g. extensive analyses for investigating the happiness paradox (friends in social networks generally seem to be happier than the considered user) or users' behavior on the online platform connected to income (Bollen et al., 2017; Preoţiuc-Pietro et al., 2015).

While the access of information seems to rise in the progressing information era, people are able to hide behind their online accounts when indicating a statement of political or societal relevant nature. Investigating online accounts offers opportunities for data scientists to understand trends and sentiments of society and to draw conclusions on relevant character traits of online platform participants. In contrast to classical clipboard surveys, analyzing online accounts may mitigate honesty biases as people are more willing to disclose information in online environments (Benartzi and Lehrer 2015). Therefore, this approach allows a valuable complementary perspective on sensitive topics (political or societal) compared to results from a questionnaire. Findings can be used to guide decisions made by policy makers in the real world as a person's personality characteristics and his/ her behavior in both, real and online world are significantly connected (Quercia et al., 2011). Findings depend on the respective chosen category of investigation. In our work we chose to investigate controversies that arise around the topic of sexuality.

Sexuality encounters openness on the one hand and refusion on the other. Discussions about sexual orientation are shaped by the history and background of conflicting parties. Modern or traditional education and religious aspects influence the opinions of the panelists. Therefore, sexual orientation is a multi-layered and wicked topic. Since the 19th century, organizations and communities have promoted a loosening of regulations against sexual orientations that are divergent to the

conventional composition of a couple as man and wife (CNN Library, 2015). Thus, they have made the discussion vivid and relevant for society. Disclosing communities that busy themselves with sexual orientation offers a better understanding of the composition of society as whole.

Our work addresses the following research question: How do machine learning techniques allow us to conclude from users' behavior and language on Twitter, to their attitudes about the LGBT movement? In order to answer this question, we first give the reader an overview about the theoretical background of our research and formulate four research hypotheses. Second, we explain our used methods in detail and reveal the results of our work. Finally, we critically discuss our findings and give an outlook for further research fields.

2 Theoretical Background/Related Research

This section will focus on discussing the fundamental definitions, that the reader will encounter through the rest of this paper. Besides that, other related work will be briefly discussed in order to show the relevance of the topic.

2.1 COINs

COINs (Collaborative Innovation Networks) are innovation networks that are often self-organized and form independently of formal organizational structures in companies or within company networks (Gloor et al., 2018).

2.2 Tribefinder

A tribe is "a network of heterogeneous persons linked by a shared passion or emotion" (Cova and Cova, 2002). The system *Tribefinder* identifies these virtual tribes. Using data on the social media platform Twitter, it analyses an individual's tweets by extracting information about key people, brands, used words and topics of his or her tweets and categorizes the user into tribes belonging to five specific tribal macro-categories: personality, alternative realities, ideologies, lifestyle, and recreation. To analyze and identify the virtual tribes the continuous stream of tweets is an important source of information, which offers a powerful setting for studying and identifying tribes of individuals (Bringay et al., 2011).

Using *Tribefinder* and the tribal vocabulary (which tribes are identified by which words or vocabulary) it learns, it is now possible to establish the tribal affiliations of every Twitter user. In practice, *Tribefinder* analyzes the individual's word usage in her or his tweets and then assigns the corresponding personality, alternative realities, ideologies, lifestyle, and recreation tribal affiliation based on the similarities with the specific tribal vocabularies.

3 Hypotheses

For the purpose of our research we formulate four hypotheses. In order to clearly predict user's attitudes towards LGBT, we need at least two groups with different attitudes that differ in their language and behavior:

H1: Two groups exist, that highly differentiate in their attitude towards the LGBT movement.
H2: These two groups use different language and reveal different honest signal characterizations.

In our work, we believe in the effectiveness of word analyses and demonstrate a bag-of-words approach:

H3: Analyzing users' words used in Twitter provides a high potential for prediction.

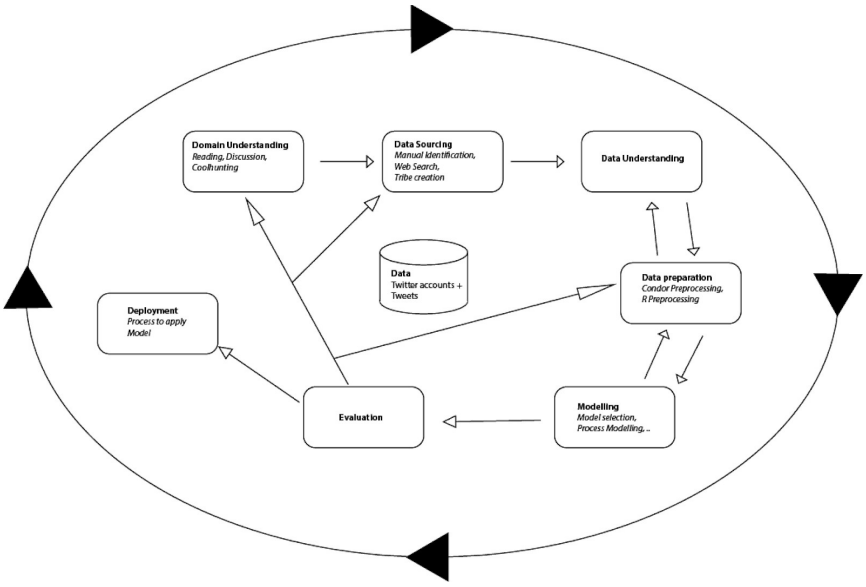
Finally, we apply our model on another tribe that consists of people who are against gun control regulations. Intuitively we consider a convergence of opinions between the Anti-LGBT tribe and the contra-gun-control tribe as more likely than between the LGBT-tribe and the contra-gun-control tribe:

H4: There are more Anti-LGBT tribe classified people in the contra-gun-control tribe than LGBT tribe classified people.

4 Methodology

To analyze large chunks of data a proper framework or guideline is required in order to find the best amount of accurate data for our project. Since Data Mining is a creative process which requires different skills and knowledge, it is very hard to tie the success of the project to the knowledge of a single team member (Wirth & Hipp, 2000). This is why we lean on the *CRISP-DM* (Cross Industry Standard Process for Data Mining) guideline which will merge our thoughts and guide us through a proper way of finding the accurate data for the development of this project (Appendix 6). Many of the required steps and processes to gather the data have been discussed and addressed in section one of this paper. The *CRISP-DM* model is divided in six phases which can interact in a cyclic pattern. The phases are categorized as: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* and *Deployment* and will be discussed in this section (Shafique & Qaiser, 2014).

Fig. 1: Slightly adjusted CRISP-DM model with additional connections between processes that promote more flexible adjustments of particular process steps after the evaluation phase



For our project we altered the first phase of the *CRISP-DM* model *Business Understanding* to *Domain Understanding*, since we are gathering and understanding information about a certain domain rather than a business venue. In our approach of *Domain Understanding* we worked around our main project task, which was to find out how different tribes with specific characteristics develop and correlate in digital networks. In order to do that, we brainstormed and gathered our ideas on which communities clash against each other the most and which ones were represented through a social media outlet such as *Twitter*. Out of this Brainstorming session we decided to analyze the correlations between LGBT and Anti-LGBT communities.

In order to discover more about the differences of the communities, we reached out to inform ourselves of the basic terminologies using *Google Scholar*, *Wikipedia* and implemented Coolhunting methods for identifying the most influential trendsetters of these characteristics. To find more information about Anti-LGBT and what it is comprised of we looked for extremist groups and websites which promote this characteristic. We also started looking for representations of these communities on *Twitter* by identifying important and common "hashtags" and popular personalities within these communities. The gathered information out of Data Understanding is discussed further in section five of this paper.

We implemented a *Data Sourcing* Phase before the *Data Understanding* phase in our model which shows an alteration from the presented CRISP-DM model. At this point we used *Condor,* a software program developed by *galaxyadvisors* which is used to measure the structure, content, sentiment and influence of social communication networks over time. *Condor* also provides visualization features which we use to better understand the data we gather. Here we used three different approaches to collect the required data, which we derived as useful from the *Domain Understanding* phase. The first approach was focused on gathering the data via the *Gal-*

axyscope-tool, *Tribe Creator* provided by *galaxyadvisors,* where a certain keyword could be used as input such as a hashtag "#" in order to filter the results by the given input.

Here the tool would provide us with users and their *Twitter* ID's, which we could use to search for friends and followers of that specific user. The second approach was to manually search *Twitter*, for specific users that would also use certain keywords, hashtags or phrases. The third approach was to use *Condor* and its *tribe-fetch* function to find certain users who also used a certain keyword. With it we obtain a list of users which we then added to the *Tribe-Creator-tool,* which was within the tool set of *galaxyadvisors*. The main focus of this phase was to create Tribes (section one), which we would later on use to create final data sets for our data mining model. The results of this phase will be thoroughly discussed in section five of this paper.

In the *Data Understanding* phase, we used the raw gathered data and implemented it in Condor 3 in order to better understand the connection between every single actor. This phase will be closely tied to the Data Preparation phase, due to the functionalities and calculations that Condor provides. Thanks to the different visualization functionalities, the user can understand how different tribes differ in structure. Besides that, *Condor* allows social network functionalities to be calculated such as the degree centrality, betweenness centrality and closeness centrality which all show the importance and position of certain actors within the network. The Results of this phase will be presented in section five of this paper.

In our *Data Preparation* steps we used different tools to properly reduce the data for its optimal and most effective use. We decided that words, their frequency and how often they appear within a certain tribe would help us to predict a certain tendency towards a tribe. This is why we needed to prepare the data in such a way, that words should be the most resonant part of the data. In order to do so we first used some of Condor 3 functionalities which calculate the six honest signals of collaboration, which are the most evident through the tweets we have collected through Twitter. The six indicators are central leadership, rotating leadership, balanced contribution, rapid response, honest language and shared context. With these signals future creativity, performance and outcomes for teams can be predicted (Gloor, 2017). Besides the six honest signals of collaboration another important way of making the words the core of our data, was to calculate the *Pennebaker Pronouns*.

Here the number of Pronouns within a tweet of every user were counted. Condor 3 has a built-in function that does so automatically and calculates the probability that a certain pronoun will appear in a tweet of the observed person (Gloor, 2017). Pennebaker discovered, that how people use pronouns, have a high predictive value (Pennebaker, 2011). After having *Condor* prepare the Data, we exported it into an R-Script which was written in the language *R.* This programming language is also an environment for statistical computing and graphics, due to its wide variety of statistics (linear, non-linear classification, classical statistical tests, classifications and more statistical calculations) it seemed the most efficient solution for our data. With the R-Script we prepared the words in such a way, that it can be representable for a Machine Learning Algorithm. The bag-of-words approach helps us in this specific task. The bag- of-words approach describes the occurrence or frequency of a word within a certain document (Brownlee, 2017). Any other information besides the words are discarded. With the number of occurrences, it is intuitive, that similar tribes will have similar words. This Phase was tightly connected with data

understanding and the modeling phase, since many iterations and changes to the data had to be made in order for it to fit our model.

In the *Modeling* phase we decided to use an online modeling tool *Rapidminer Studio. Rapidminer Studio* is a visual workflow designer, which helps develop prototypes for predictive models. Its GUI (graphical user interface) and provided documentation leads the user through the whole process of modeling and provides further information about every function, algorithm or component that is used (Rapidminer, 2019). We integrated our prepared data into the tool and applied all predictive Machine Learning Algorithms available in the toolset of *Rapidminer Studio*. best practice in *Rapidminer*, is to cross validate, training-set and test-set which provide values such as accuracy and recall to better select the best decision. After the *Modeling* phase the *Evaluation* of the model is required. Here all results of the algorithms will be taken under consideration. Our decision will be mainly made by the highest accuracy provided by models which were calculated with different Machine Learning Algorithms. Accuracy is calculated by the percent- age of correct predictions over the total of examples we fed the model and by correct prediction we mean that the value of prediction corresponds to the label attribute we specifically picked in the Modeling phase and applied to the Rapidminer model. The results of the Modeling and Evaluation phase is discussed in section five.

It is important for us to develop a model, which allows to be used for two scenarios. Firstly, for predicting a certain tribe within another tribe, and secondly, for predicting a user's tendency towards one tribe or another depending on his/her tweets. In the Deployment phase we prepared the model in such a way, that it is accessible for every example and data set. This is achieved by providing a documentation of how to use the model and where to introduce the example data set. To conclude our methods used during this project it is important to understand the iterative and cyclic nature as seen in Figure 1. Every phase can be altered in order to adjust the final data set to provide the best possible outcome of the intended predictive model. Within the Deployment phase the sub-phase Demonstration takes its place. A finished model used with real time data tests its potential prediction.

5 Results

In this section, we will present the results structured by the phases of our *CRISP-DM* adjustment (see section four). We worked iteratively during the project making use of the loops the methodology provides. In order to provide clear overview, we will only present the results of the last iteration of the respective phases here.

Domain Understanding

Sexualities split up into several groups. There is heterosexuality which can be considered the most traditional and popular sexuality and describes the sexual preference for the respectively other gender. Besides there are rather alternative sexual preferences such as homosexuality, bisexuality, transsexuality and others. Finally, most alternative sexualities sum up in LGBT movement. Therefore, we choose this group as a major tribe for our considerations. LGBT stands for lesbian, gay, bisexual and transgender. Moreover, variants such as LGBTQ, LGBTQ+, LGBTQI+ exist, which is also reflected in hashtag usage. All these terms usually refer to the same community and the basic idea that open-minded- ness towards sexuality is important and one should tolerate all sexual minorities. As a result of our Coolhunting we

identified that LGBT is the most common hashtag and community that is referred to. Therefore, we defined our LGBT tribe as people who openly support lesbian, gay, bisexual or trans. In order to get contrasting training data for our final model, we consider people who are significantly different from LGBT supporters. Therefore, we looked at people who are opposed to the LGBT movement. Typically related keywords in literature are homophobia and transphobia. In the course of our explorative research on Twitter, we identified a few potential subtribes regarding these attitudes. The spectrum reaches users on Twitter who are opposed to gay marriage to users who express in their tweets that alternative sexualities are diseases, that need to be cured, and users who verbally attack LGBT communities on Twitter in a disrespectful way. To include these different phenomena, we generally defined our Anti-LGBT tribe as people who are opposed to LGBT as sexual orientations.

Data Sourcing

Currently, the V1 LGBT tribe collected in *Tribefinder* contains 168 members who actively use Twitter. V1 Anti-LGBT consists of 119 members. The tribe-fetch with Condor resulted in two datasets of network (Twitter) data containing a total of more than 20,000 actors (users) and 480,000 links (tweets) including all the tribe members and their respective social networks on Twitter.

Data Understanding

This stage was highly interrelated with the consecutive data preparation stage (see section 4. Methodology). Therefore, we include results regarding features that were actually generated by the later data preparation stage. Apart from Condor generated features and visualizations, we look at the tribe member datasets resulted from data preparation including bag-of-words features. A look on the network graphs in Condor gives a first insight into the different tribes. The graphs depicted in Appendix 1. display all actors and links to the respective tribes in their surrounding network. The node color yellow highlights tribe members, the node size scales with the betweenness centrality measure. Whereas both networks seem quite strongly connected, the LGBT network looks a bit dominant in this respect. Tribe members in LGBT are more often strongly connected and further in the middle of the graph. It is striking that the LGBT network is showing more non-tribe members that are quite central as well. In contrast, the Anti-LGBT network shows that few tribe members are very central in the network (big yellow nodes) but there are few central nodes of other tribes in the network. This could likely mean, that the Anti-LGBT community is more isolated from and less connected with non-tribe related important people. Moreover, there are mainly very central leaders and many non-central followers in the Anti-LGBT network. The tendency of centrality in the Anti-LGBT tribe versus collaborative decentrality in the LGBT network is also reflected in the t-tests results and boxplots (Appendix 2). The median LGBT tribe member has a lower betweenness and degree centrality than the median Anti-LGBT tribe member. Betweenness Centrality Oscillation, however, is relatively dominated by the LGBT tribe. The word clouds generated with Condor (Appendix 3) give us a good feeling for the language use of our two different tribes. The size of terms depicts the relative frequency of terms in tweets. The color indicates the detected sentiment ranging from negative (red) to positive (green). It is obvious that the LGBT tribe has an overall more positive sentiment than the Anti-LGBT tribe (which is also confirmed by t-test results). Regarding the content, we find that, Anti-LGBT

tribe members significantly more often use political terms (e.g. wall, bill, nation, democrats, senate, Obama, Trump) and religious terms (e.g. christian, god), whereas the language of LGBT tribe members is rather dominated by social terms (e.g. community, friends, family, people, today) and LGBT related terms (e.g. trans, love, pride, gay, lgbt, transgender, person). Regarding Pennebaker Pronouns, a look on the word usage distributions suggests that LBGT community members tend to use pronouns in a self-related way, if the pronoun is personal, while the Anti-LGBT community tends to use more non-personal pronouns - or personal pronouns linking to other people. In particular, the t-tests validate that "my", "me" and "it" are significantly more often used in the LGBT tribe. Anti-LGBT tribe members on the other hand signficantly more often used the pronoun "the". Moreover, they use the pronouns "his", "they" and "that", coming as bag-of-words features, significantly more frequently (Appendix 7). These findings are also reflected in the weights of the final model's features, suggesting that pronouns features do well on contributing to the predictability of tribe membership (Appendix 8 & 9).

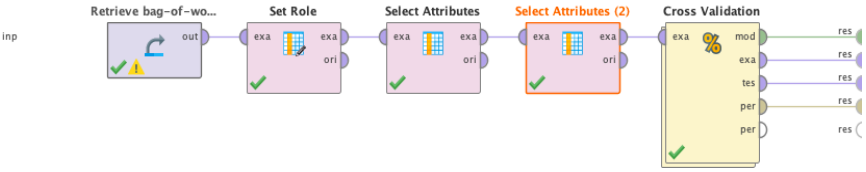
Data Preparation

As a result of data preparation, the final training and evaluation dataset is a combined dataset of all our tribe members with 134 features plus our target variable, the tribe name. 49 of the attributes come from the fetched actor data itself as well as metrics that are calculated from the actors' network by Condor (six honest signals and Pennebaker pronoun frequencies). Moreover, there are 85 attributes that are generated from the aggregated link data by means of our bag-of-words processing.
It should be noted that the final training dataset consists of 115 Anti-LGBT entries and 111 LGBT entries due to filters in the process such as the filter in Condor that removes actors with much less activity for meaningful metrics.
Within the final modeling process in *Rapidminer* (see next paragraphs), we finally deselected some of the features. Features with too many missing values as well as identity-like attributes, such as names. This resulted in a final training dataset with 226 rows and 105 columns (features).

Modeling

Our model classifies a Twitter user as a LGBT (or Anti-LGBT) tribe member, given the entity including all its 104 features. Based on our Evaluation we choose a Generalized Linear Model, a machine learning model for classification problems such as ours. We trained the model using the *Rapidminer* process (depicted in Figure 2). The process consists of four data processing steps: data retrieval from the imported training data, selection of the target variable (Tribe), the final selection of features to be used for training, and the training and testing within a cross validation (see Figure 3).

Fig. 2: Process in Rapidminer



Evaluation

The cross-validation of different feature sets and machine learning algorithms revealed best results for our final model, which utilizes a Generalized Linear Model and 104 features. The final model's evaluation results are depicted in Figure 3. The performance can be summarized with 77,43 % accuracy. The model performs slightly more precise on Anti-LGBT predictions (precision: 79,63 % vs. 75,42 %) and slightly better recalls true LGBTs (80,18 % recall vs. 74,78 %). In other words, if an actor is classified as Anti-LGBT it is more likely to be correct, and if an actor is LGBT it is likely that he correctly gets detected as such, than it is respectively to correctly classify a LGBT or detect every Anti-LGBT.

Fig. 3: Final model confusion matrix and accuracy

accuracy: 77.47% +/- 6.78% (micro average: 77.43%)

	true V1 Anti-LGBT_483569...	true V1 LGBT_847ec453ee5...	class precision
pred. V1 Anti-LGBT_48356...	86	22	79.63%
pred. V1 LGBT_847ec453ee...	29	89	75.42%
class recall	74.78%	80.18%	

In order to decide on a specific algorithm, we tested six different machine learning methods with *Rapidminer* Auto Model. It revealed that Naive Bayes and Generalized Linear Model performed best (see Appendix 5.). A follow-up analysis in the custom *Rapidminer* process proved the Generalized Linear Model performs best for our final attribute selection cross validation. Our evaluation also demonstrates the improvement caused by the inclusion of bag-of-words features. The cross-validation robustly shows that there is an improvement around 8 % (77,43 % instead of 69,03 %, Figure 4.).

Fig. 4: Model confusion matrix and accuracy without bag-of-words features.

accuracy: 68.99% +/- 10.13% (micro average: 69.03%)

	true V1 Anti-LGBT_483569...	true V1 LGBT_847ec453ee5...	class precision
pred. V1 Anti-LGBT_48356...	82	37	68.91%
pred. V1 LGBT_847ec453ee...	33	74	69.16%
class recall	71.30%	66.67%	

The final configuration of bag-of-words specifies the maximal allowed sparsity parameter as 0.985. Words from messages are *stemmed* and *stop words* are not removed. Regarding this configuration, we did not evaluate all possible configurations, but took a look at different configurations within a reasonable range. Better results were reached with a higher maximal allowed sparsity level. However, we limited the allowed word sparsity at some point to keep the number of attributes relatively low. Stemming words and not removing words were proved to be dominant over all other combinations of these Booleans in terms of resulting model accuracy.

Deployment

One goal of ours evolved to be a deployable solution that allows model application. To apply the model on new entities, we fetched the *Twitter* accounts of three single actors, namely Eminem, Donald Trump and Peter Gloor. Moreover, we fetched another tribe, the *contra-gun-control* tribe. The application of our model yields the following results. Members of the *contra-gun-control* tribe are people who are supposed to like guns. According to our model they are mainly identified as Anti-LGBT (LGBT: 0.286 vs. Anti-LGBT 0.714). These single actors give us a good range of results. Donald Trump (*@realdonaldtrump*) is identified as an Anti-LGBT with a confidence of 97,6 %. Marshall Mathers (*@Eminem*) is identified as an Anti-LGBT with a confidence of 73 %. Finally, Peter Gloor (*@pgloor*) is identified as an LGBT with a confidence of 60,4 %.

6 Discussion

Looking at our results allows us to draw conclusions to bolster our hypotheses. Results from the domain understanding indicate, that there are at least two different groups, that highly differentiate in their attitudes towards the LGBT movement (H1). Our LGBT and Anti-LGBT tribes represent the two different groups that are attuned in either positive or negative way towards the LGBT movement. Positions inside those groups can be (especially in the Anti-LGBT group) versatile in its level of aversion or affection. During the data sourcing we built two decent tribes by extensively manual inspecting every Twitter account for its veracity of attitude, that is desired for the respective tribe. Therefore, researchers can use those data from our tribes for further analyses as a solid fundament for their work. In the stage of data understanding we show that language and behavior differs between the members of the two tribes (H2). This manifests for example in sentiment and centrality measures and also in the word use of the tribe members. After a proper data preparation, results of our evaluation phase indicate a high prediction potential for analyzing the used words by users (H3). Including bag-of-words features shows an improvement of around 8 % in cross-validation. Interestingly, more general terms such as pronouns and conjunctions are shown to be more meaningful for our prediction value then more goal content specific words. Demonstration inside our deployment phase indicates that there are LGBT tribe classified people in the contra-gun-control tribe. However, the proportion of Anti-LGBT classified people in the contra-gun-control tribe is significantly bigger. Therefore, our hypothesis H4 can be obtained.

While we achieve strong results that are intended for satisfying support of our hypotheses, several limitations have to be taken into account. A rising quantity of data impedes the process of machine algorithm calculations and the preparation of sound prediction models, which leads us to limit the data quantity.

Nevertheless, concentrating on a limited quantity of data enables quality improvements like aiming at manually minimizing poor data as fake accounts and fake tweets, even if we do not emphasize nor quantify this procedure further. To improve the quality of our predictions we mainly focused on accuracy. We do not minimize the complexity of used features as we want to ensure a maximal accuracy irrespective of performance efficiency aspects. We do not investigate possible trade-

off effects on accuracy and performance by limiting or adding different prediction features. In consideration of practicality aspects, we also consider the option of developing a more user-friendly IT-artifact as a proficient way for suitable applicability. Our present approach is more of a "do-it-yourself" one. Furthermore, our model is strongly attached to a certain domain. While it does well in the LGBT context there is no proof that our procedure performs on a same level in other domains of use. Interested scientists could aim at diminishing the above-mentioned limitations by elaborating on our research in further investigations. Also we propose to expand the domain field of application to other areas. Applying our model to other tribes, like for example religion tribes, can provide insights into effects from tribe affiliations (like religious affiliation) on attitudes towards the LGBT movement. Our model also provides opportunities in the field of tailored marketing. Identifying a person's attitudes about a certain field can lay the foundation to create customized advertisements in a next step. Though, moral issues should be taken into consideration, because this approach is likely to be manipulating. All in all, our work offers various insights into machine learning techniques for identifying attitudes from Twitter language and behavior plus a well-applicable model for the domain of the LGBT movement. While there is potential for further investigation, all of our previous formulated hypothesis can be obtained.

7 Conclusion, Outlook & Limitations

In our work we show how machine learning techniques allow us to conclude from users' behavior and language on Twitter their attitudes about the LGBT movement using an adjusted procedure of the *CRISP-DM* process. By identifying two groups of contrary attitudes towards LGBT, we create two tribes by using the tool Tribefinder. We show that language and behavior of users in the respective tribes differ. Furthermore, we identify word analyses as valuable mean of prediction. Thereby, specific terms are not as decisive as general ones like pronouns or conjunctions. Applying our model on the data set of the contra-gun-control tribe reveals that the proportion of Anti-LGBT classified people in the contra-gun-control tribe is significantly bigger than LGBT classified people. The application of our prediction model on single Twitter accounts to identify a single users' attitudes towards the LGBT movement gives us comprehensible results. Further research could investigate how higher data quantities affect the model's quality. Furthermore, investigations could aim at applying our model in different domains than the LGBT movement.

References

Apelt M (2010) Forschungsthema: Militär: Militärische Organisationen im Spannungsfeld von Krieg, Gesellschaft und soldatischen Subjekten, 1. Aufl. VS Verl. für Sozialwiss, Wiesbaden.

Benartzi S, Lehrer J (2015) The smarter screen: Surprising ways to influence and improve online behavior. Portfolio/Penguin, New York, New York.

Bollen J, Gonçalves B, van de Leemput I, Ruan G (2017) The happiness paradox: your friends are happier than you. EPJ Data Sci. 6:497. doi: 10.1140/epjds/s13688-017-0100-1.

Bringay S, Béchet N, Bouillot F, Poncelet P, Roche M, Teisseire M (2011) Towards an online analysis of tweets processing. Database and Expert Systems Applications. Springer, Heidelberg, Berlin.

Brownlee, J. (2017, Oktober 8). A Gentle Introduction to the Bag-of-Words Model. Abgerufen 14. März 2019, von https://machinelearningmastery.com/gentle-introduction-bag-words-model/.

Cambridge Dictionary "Tribe". https://dictionary.cambridge.org/de/worterbuch/englisch/tribe. Accessed 8 March 2019.

CNN (2019) “The next LGBT cause: gun control”. cnn.com (2016). https://edition.cnn.com/2016/06/24/health/lgbt-gun-activism/index.html (accessed 10.03.2019).

CNN Library (2015) LGBT Rights Milestones Fast Facts. https://edition.cnn.com/2015/06/19/us/lgbt-rights-milestones-fast-facts/index.html. Accessed 8 March 2019.

Cova B, Cova V (2002) Tribal marketing: The tribalisation of society and its impact on the conduct of marketing European Journal of Marketing 36:595-620.

Cross, R. L., Cross, R. L., & Parker, A. (2004). The hidden power of social networks: Understanding how work really gets done in organizations. Harvard Business Press.

De Oliveira, J. M., & Gloor, P. A. (2018). GalaxyScope: Finding the “Truth of Tribes” on Social Media. In Collaborative Innovation Networks (pp. 153-164). Springer, Cham.

Eckerson, W. W., Hanlon, N., & Barquin, R. (2000). DIRECTOR OF EDUCATION AND RESEARCH, 5(4), 15.

Galaxyscope.galaxyadvisors.com https://galaxyscope.galaxyadvisors.com/tribe/donaldtrump, accessed 11.03.2019.

Gloor, P. A. (2017). Sociometrics and Human Relationships. Abgerufen von https://www.emeraldinsight.com/doi/abs/10.1108/978-1-78714-112-420171027.

Gloor, P., Colladon, A. F., de Oliveira, J. M., & Rovelli, P. (2018) Identifying Tribes on Twitter through Shared Context.

Merriam-Webster Dictionary "Tribalism". https://www.merriam-webster.com/dictionary/tribalism. Accessed 8 March 2019.

Oliveira JMD, Gloor PA (2017) GalaxyScope – Finding the “Truth of Tribes” on Social Media.

Preoţiuc-Pietro D, Volkova S, Lampos V, Bachrach Y, Aletras N (2015) Studying User Income

through Language, Behaviour and Affect in Social Media. PLoS ONE 10:e0138717. doi: 10.1371/journal.pone.0138717.

Pennebaker, J. W. (2011). The secret life of pronouns. New Scientist, 211(2828), 42–45. https://doi.org/10.1016/S0262-4079(11)62167-2.

Pritchard D (2000) Tribal Participation and Solidarity in Fifth-Century Athens: A Summary. Ancient History:104–118.

Quercia D, Kosinski M, Stillwell D, Crowcroft J (2011) Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In: IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT), 2011 and 2011 IEEE Third International Conference on Social Computing (SocialCom): 9 - 11 Oct. 2011, Boston, Massachusetts, USA ; proceedings ; [including workshop papers. IEEE, Piscataway, NJ, pp 180–185.

Rapidminer. (2019). RapidMiner Studio - RapidMiner Documentation. Abgerufen 14. März 2019, von https://docs.rapidminer.com/latest/studio/.

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA), 12(1), 217–222.

Statista (2016) statista Number of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2018 (in millions). https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/. Accessed 8 March 2019.

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In 2012 11th International Conference on Machine Learning and Applications(Vol. 2, pp. 386-393). IEEE.

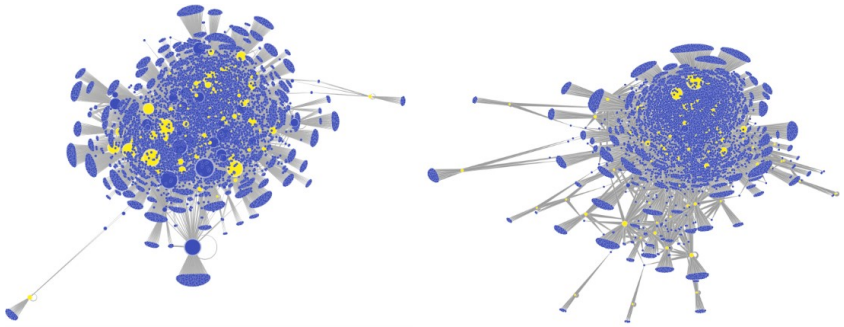
Terpstra, T., De Vries, A., Stronkman, R., & Paradies, G. L. (2012). Towards a realtime Twitter analysis during crises for operational crisis management (pp. 1-9). Burnaby: Simon Fraser University.

Twitter, Inc. Twitter für Unternehmen. https://business.twitter.com/de.html. Accessed 14 March 2019.

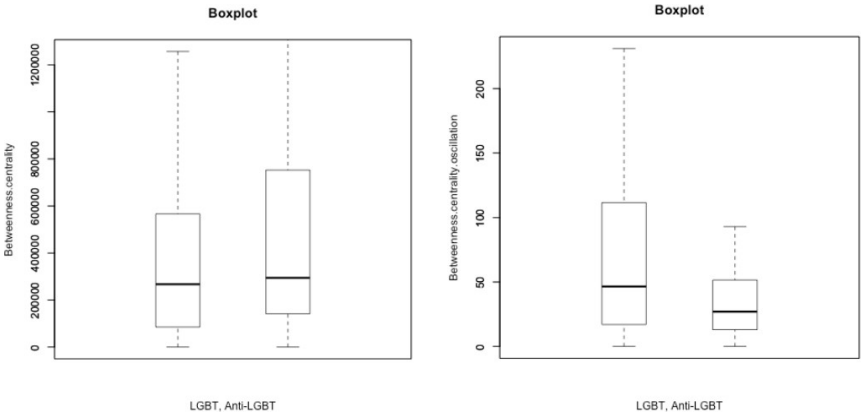
Wikipedia.org https://en.wikipedia.org/wiki/LGBT accessed 06.03.2019.

Wirth, R., & Hipp, J. (2000). CiteSeerX — CRISP-DM: Towards a standard process model for data mining. Abgerufen 9. März 2019, von http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.198.5133.

Appendix

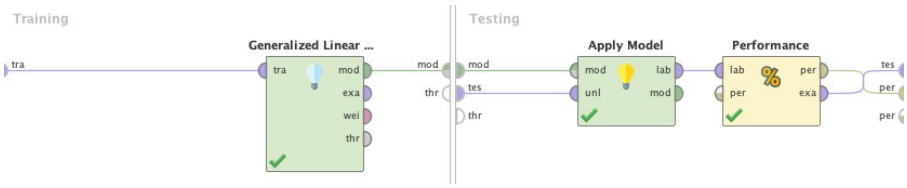


Appendix 1: Tribal network graphs of LGBT (left) and Anti-LGBT (right) tribes.

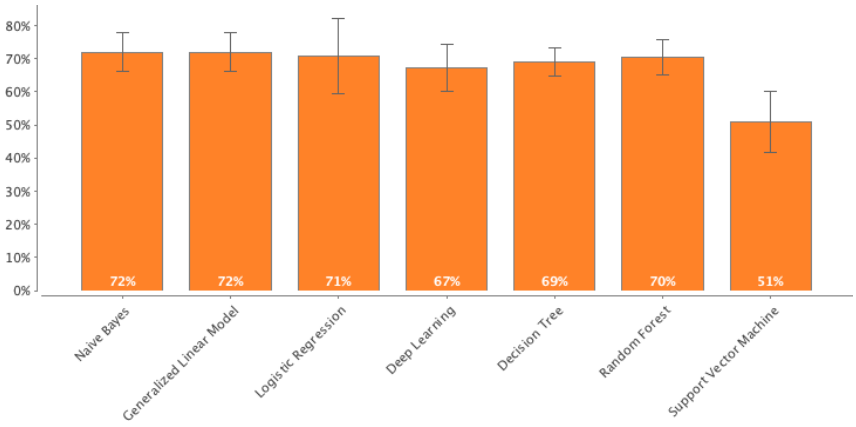


Appendix 2: Boxplots of betweenness centrality and betweenness centrality oscillation.

Appendix 3: Tribal word clouds of LGBT (left) and Anti-LGBT (right).



Appendix 4: Subprocess of cross-validation.



Appendix 5: Performance comparison of different classification methods with Rapidminer Auto Model

Steps:	Tools:	Methods:
Domain understanding	Twitter, Web-search	Reading, Discussion, Coolhunting
Data sourcing	Condor, Tribe Creator	Manual identification, Web search, Tribe creation
Data understanding	Condor, R-Studio	Statistical analysis methods
Data preparation	Condor, R-Studio, RapidMiner	Condor Preprocessing, R preprocessing
Modelling	RapidMiner	Model selection, process modeling
Evaluation	RapidMiner	Data Science evaluation measures
Deployment	Condor, R-Studio, RapidMiner	

Appendix 6: Overview of process steps by Software/ methods used

feature	p-value	lgbt_mean	anti_lgbt_mean
wall	2,73E+04	0.00181983330106428	0.0196835902088849
his	1,49E+05	0.0186030868611337	0.0385758966402175
presid	1,07E+06	0.0084632145513462	0.0266512267771265
democrat	1,48E+06	0.00488378672481424	0.023664798539182
will	8,01E+06	0.033411350465989	0.0581733192293614
they	2,96E+07	0.0285739054133454	0.0476930103709141
realdonaldtrump	4,40E+08	0.0179948844619235	0.0601055732114953
the	5,11E+09	0.418356562837061	0.522607922217094
peopl	0.000100533358739003	0.0525751221937414	0.0342322675309065
frequency_my	0.000135047097344263	0.00243479387222333	0.00126774235489974
frequency_me	0.000224745296962141	0.00199155660170316	0.00106036260005562
what	0.000458726842254082	0.0357170358214091	0.0495760177422029
should	0.000743342427618055	0.0140121480236884	0.020119864476807
Degree.centrality	0.00139158365662965	127.357.142.857.143	232.394.957.983.193
love	0.00172078485229054	0.035157306334108	0.0231379494996043
not	0.00179916978065713	0.0529034090616066	0.0682407490824305
statuses_count	0.00267150224494203	86.450.625	305.316.302.521.008
trump	0.00394782211251797	0.0198831201761341	0.0355391699705567
avg.sentiment	0.0055189834191805	0.517823002905533	0.479268834259697
whi	0.00726373382483415	0.016467524192345	0.021793280395551
look	0.00731389713599833	0.0233503410138929	0.0166119392092321
Betweenness.centrality.oscillation	0.00773359301712264	729.910.714.285.714	474.957.983.193.277
say	0.0106401217215972	0.02054332887485	0.025422998447025
get	0.0125406533127217	0.0337990085701445	0.0410704625509712
when	0.0131384930931877	0.0263026697629361	0.0334699716108629
know	0.01788637672863	0.0210117758058814	0.026167714922749
right	0.0189868852337561	0.0254718642426327	0.019000337684749
has	0.0215273577825814	0.0297930975778801	0.0368333844412554
avg.emotionality	0.0268918462896586	0.260084620104577	0.266640160515609
messages.sent	0.0289407993099403	47.872.972.972.973	727.822.033.898.305
this	0.0305944896060258	0.116576111161055	0.10169841927249
would	0.0345370635668361	0.0151652754556052	0.0199027446397517
frequency_it	0.0355062768340602	0.00322191758683079	0.00261660806975636
frequency_the	0.0396780822093489	0.0118753276060678	0.0134107995952259
Betweenness.centrality	0.0403804804570557	44.839.319.467.418	678.124.189.817.623
want	0.0431226000358793	0.0194807087847524	0.0237834510715019
Messages.sent	0.047790051320594	798.803.571.428.571	115.948.739.495.798
that	0.0500117914408883	0.0952595064308129	0.11075080516433

Appendix 7: T-tests results sorted by p-value, cut at p<=0,05

Attribute	Coefficient	Std. Coefficient ↑
wall	-24.037	-0.516
will	-12.100	-0.443
they	-13.207	-0.385
what	-10.200	-0.316
his	-11.678	-0.308
when	-11.418	-0.256
presid	-8.830	-0.221
the	-0.883	-0.177
democrat	-6.685	-0.173
good	-9.214	-0.151
Intercept	0.676	-0.126
whi	-7.068	-0.108
avg.emotionality	-2.956	-0.068
Degree.centrality	-0.000	-0.053
follow	-1.276	-0.049
you	-0.430	-0.047
know	-1.715	-0.029
Betweenness.centrality	-0.000	-0.023
all	-0.568	-0.015
not	-0.226	-0.009
but	-0.029	-0.001

Appendix 8: Features which the generalized linear model attributes to Anti-LGBT.

Attribute	Coefficient	Std. Coefficient ↓
peopl	13.720	0.490
are	5.263	0.244
right	11.311	0.236
have	6.744	0.233
look	11.171	0.212
this	2.583	0.136
frequency_me	72.970	0.134
Betweenness.centrality.oscillation	0.002	0.126
who	4.654	0.121
frequency_my	52.673	0.119
trump	2.552	0.106
new	2.732	0.081
love	2.467	0.073
avg.sentiment	0.592	0.063
our	0.920	0.058
frequency_it	26.302	0.055
vote	1.655	0.053
frequency_have	49.927	0.051
Contribution.index	0.140	0.042
frequency_was	17.980	0.023
friends_count	0.000	0.007
there	0.349	0.006

Appendix 9: Features which the generalized linear model attributes to LGBT members.