Department of Information Engineering and Computer Science

Bachelor's Degree in
Computer Science

Final Dissertation

Analysis of emotions in Wikipedia discussions

Supervisor	Co-supervisor	Student
Alberto Montresor	Cristian Consonni	Nicola Toscan
	David Laniado	

Academic year 2020/2021

Acknowledgements

This project was possible only thanks to people truly interested in the topic, who guided and helped me in the past months. A special thanks goes to Cristian Consonni and David Laniado, who followed my progress with interest and are always available to help. They always went a step further than what they were asked. I also want to thanks all the other students in our team, which whom I collaborated in the most pleasant way. Finally, I want to thanks professor Alberto Montresor, who introduced me to this project and was always available and interested in my progress.

Contents

Abstract **3**

Sommario **4**

1 Introduction **5**
1.1 Our project . . . 5
1.2 My contribution . . . 6

2 Background **7**
2.1 Wikipedia . . . 7
2.1.1 Wikipedia Users . . . 7
2.1.2 Wikipedia talk pages . . . 8
2.2 WikiConv Dataset . . . 8
2.3 Emotional analysis . . . 9
2.3.1 EmoLex . . . 9
2.3.2 LIWC . . . 10

3 Methods **11**
3.1 Categorizing users . . . 11
3.1.1 Users' gender . . . 11
3.2 Lookup tables . . . 12
3.3 Sorting the WikiConv dataset . . . 13
3.4 Minifying the WikiConv dataset . . . 14
3.5 Analyzing the WikiConv dataset . . . 15
3.6 Metrics . . . 16

4 Results and Discussions **17**
4.1 Users' gender . . . 17
4.1.1 UserBoxes . . . 17
4.1.2 Profile settings . . . 18
4.2 Emotions . . . 19
4.2.1 Word Clouds . . . 19
4.2.2 Users . . . 20
4.2.3 Pages . . . 22

5 Infrastructure **24**
5.1 Storing the dataset . . . 24
5.2 Storing our results . . . 24
5.3 Public API . . . 24
5.4 Web Application . . . 25
5.5 Deploy . . . 26

6 Conclusions **27**

Bibliografia **27**

Abstract

Wikipedia is a free, open-source encyclopedia. Its content is written by a community of volunteer editors. It has grown through the years, becoming the biggest encyclopedia ever made. Wikipedia is composed of articles, which collects the actual encyclopedia and talk pages, where editors can interact which each other.

This project is a part of a larger project that wants to understand when and why Wikipedia editors stop to edit. As all online community, there are several factors that lead users to join or leave this project. Our goal is to discover which particular events in a editor life-cycle leads to quit contribution to Wikipedia. This information can be used to better the experience for all users, and increase the size of the community.

My focus was to analyze emotions expressed by editors in Wikipedia talk pages posts. We tried to understand what emotions and sentiments users express and receive while interacting with each other. We used used "NRC Word-Emotion Association Lexicon" to map a set of words to the emotions and sentiments they express. The data I generated, joined to the results from other team members, will be used to analyze users' life cycles.

All users interactions in Wikipedia talk pages were collected in the WikiConv dataset, including modifications and deletions of posts. We used this data to count emotions expressed by different users and on different pages. In particular, we analyzed users for each month since the day a user joined Wikipedia. To avoid privacy violations we grouped users by gender and roles in Wikipedia, and no result on a specific user was ever published.

Defining a user group was an aspect of particular relevance in this research. We used two different approaches to define their gender, UserBoxes on users' pages and the gender specified by each user in the Wikipedia settings. The roles were retrieved from a dataset generated by another team member.

All the results we computed were loaded into a Postgres database and are publicly available through a GraphQL endpoint. A web application was also developed to show simple charts with different metrics.

Sommario

Wikipedia è un'enciclopedia libera e open source. Il suo contenuto è scritto da una comunità di editor volontari. È cresciuta di molto negli anni, diventando la più grande enciclopedia mai realizzata. Wikipedia è composta da articoli, che raccolgono l'enciclopedia vera e propria e le pagine di discussione, in cui gli editori possono interagire tra loro.

Questo progetto fa parte di un progetto più ampio che vuole capire quando e perché gli editor di Wikipedia smettono di scrivere articoli. Come tutte le comunità online, ci sono diversi fattori che portano gli utenti ad aderire o ad abbandonare questo progetto. Il nostro obiettivo è scoprire quali particolari eventi nel ciclo di vita di un editore portano a interrompere il contributo a Wikipedia. Queste informazioni possono essere utilizzate per migliorare l'esperienza di tutti gli utenti e aumentare le dimensioni della community.

Il mio obiettivo era analizzare le emozioni espresse dagli editori nei post delle pagine di discussione di Wikipedia. Abbiamo cercato di capire quali emozioni e sentimenti gli utenti esprimono e ricevono mentre interagiscono tra loro. Abbiamo usato il "NRC Word-Emotion Association Lexicon" per collegare un insieme di parole alle emozioni e ai sentimenti che esprimono. I dati che ho generato, uniti ai risultati degli altri membri del team, verranno utilizzati per analizzare i cicli di vita degli utenti.

Tutte le interazioni degli utenti nelle pagine di discussione di Wikipedia sono state raccolte nel set di dati WikiConv, incluse le modifiche e l'eliminazione dei post. Abbiamo usato questi dati per contare le emozioni espresse da utenti diversi e su pagine diverse. In particolare, abbiamo analizzato gli utenti per ogni mese dal giorno in cui un utente si è iscritto a Wikipedia. Per evitare violazioni della privacy abbiamo raggruppato gli utenti per genere e ruoli in Wikipedia e non abbiamo mai pubblicato alcun risultato su utenti specifici.

La definizione di un gruppo di utenti è stato un aspetto di particolare rilevanza in questa ricerca. Abbiamo utilizzato due approcci diversi per definire il loro genere, le UserBox sulle pagine degli utenti e il genere specificato da ciascun utente nelle impostazioni di Wikipedia. I ruoli sono stati recuperati da un set di dati generato da un altro membro del team.

Tutti i risultati che abbiamo calcolato sono stati caricati in un database Postgres e sono pubblicamente disponibili tramite un endpoint GraphQL. È stata inoltre sviluppata un'applicazione web per mostrare grafici semplici con metriche diverse.

1 Introduction

"Wikimedia is a global movement whose mission is to bring free educational content to the world". [1]

Under this movement, many projects have risen, first among all Wikipedia, the most read encyclopedia in history and one of the top 15 most popular sites in the world.[2] It is free, multilingual, and maintained by a community of volunteers through a model of open collaboration. The project carries no advertisements and is hosted by the Wikimedia Foundation, an American non-profit organization funded mainly through user donations.

Wikipedia is considered the most successful open and free project, not controlled by any company, where anybody can freely collaborate. Since its initial release in 2001, Wikipedia has continuously grown and is now counting 233 million pages in more than three hundred different languages. In a month, the website received roughly seven billion visit and about 9 million edits.

Opposed to private companies with a community of similar size, the Wikimedia Foundation makes its data public and easy accessible to anybody interested.[3] Each month, a new dump of Wikipedia is created and published. This large amount of data can be extremely useful to data scientists willing to work on it. A community of millions can be studied and analyzed in a way never thought possible just a few years ago.

Wikipedia can continue to work only thanks to volunteers, who spend their time helping this project to grow. It is of vital importance to keep this community united and motivated, understand its actions and motivations and figure out why certain actions are taken and what is their effect on the community.

1.1 Our project

Our team worked on a series of related projects called "Community Health Metrics: Understanding Editor Drop-off" in collaboration with the Wikimedia Foundation, Eurecat, and the University of Trento. [4]

As stated in the project idea: "... we plan to carry out an extensive study of the editor's lifecycle. Special attention will be devoted to underrepresented groups according to social dimensions such as gender or geographic provenance. We will extend state of the art metrics to analyze different language editions, combining a computational approach with qualitative inspection of the findings, involving expert editors from the communities for this task when necessary. This will increase our understanding of factors that are important for community health in Wikipedia, and it will result in explainable metrics that can be applied to signal early if a page or set of pages are undergoing detrimental dynamics".

We currently lack the knowledge to understand and prevent drop-off for experienced editors. Our goal is to get a clearer picture of the phenomenon through the analysis of editors' life cycles. We want to understand the dynamics and the factor associated with editor drop-off, increasing awareness about m.the community health. Some research on this topic has already been done by members of our team[4].

We want to generate metrics and indicators about Wikipedia pages and groups of users characterized by gender, country, and native language. The results will be made available to the community through a dashboard.

[1] `https://www.wikimedia.org/`

[2] `https://wikipedia.org/`

[3] `https://dumps.wikimedia.org/`

[4] `https://meta.wikimedia.org/wiki/Grants:Project/Eurecat/Community_Health_Metrics:_Understanding_Editor_Drop-off`

1.2 My contribution

My focus was the emotional analysis of users and talk pages of Wikipedia. We tried to identify factors associated with editors leaving the project.

As in all human interactions, emotions play a crucial role in our decisions. Emotions can influence our actions, and actions can influence our emotions. Understanding this correlation can unveil recurring patterns in users' interaction to help identify problems in a community and find solutions.

I have analyzed words used by editors in a different context, tried to understand their emotional weight in a discussion and how it reflects on a user. Through all user interaction with the community, I have reconstructed its life cycle, in particular, which action led to which emotions and vice-versa. This topic was already covered in different studies, using different lexicon dictionaries and smaller datasets in 2012 [3] and 2014 [2].

I also contributed to the categorization of users based on their gender and the development of the dashboard.

2 Background

To fully understand the extent of this project it is important to understand how Wikipedia works. In this chapter I will provide an introduction to Wikipedia, its community and some relevant internal dynamics. I will also introduce some of the dataset and libraries we used.

2.1 Wikipedia

Wikipedia is a free, multilingual online encyclopedia written and maintained by a community of volunteer contributors through a model of open collaboration. Wikipedia contents are divided into pages, and pages are written by editors.

There exist 323 language editions of Wikipedia. Each edition has different pages and slightly different rules. The most popular edition is the English one, which receives 48% of Wikipedia cumulative traffic.

The total number of pages in all languages is 233 million; 57 millions of these are articles and 6 million of those are in English. English articles, not only are the most common but are also considered the most accurate, solid, and exhaustive. It is a good approximation to say that 2/3 of Wikipedia is written in English.

Wikipedia pages are grouped into namespaces, which separates data into core sets, differentiating content for public viewing and content intended for the editing community. The more relevant namespaces are "Articles" that contains the actual encyclopedia; "Talk", where editor discuss improvement to articles; "User", that contains interpersonal discussion and personal content; "User Talk", where user can send messages to each other.

All Wikipedia pages are written in *Wikitext*, a special markup language used to graphically represent a page structure. All content is backed up and saved in a dump file every month and made publicly available. The dump contains all modifications made to every Wikipedia page at all times.

2.1.1 Wikipedia Users

Wikipedia pages can be edited by registered or anonymously users. As of June 2021, there are 96 million accounts, of which about 310 thousand have made at least one edit in the last thirty days and are considered active.

Wikipedia users have different abilities to perform specific actions based on the user groups they are in. The most popular user groups and the ones of relevance for this study are:

- **Unregistered**: Users who are not logged in and are identified by their IP Address.
- **Registered**: Users that are logged in with an email and have obtained a username.
- **Autoconfirmed**: User that are at least four days old and have made at least 10 edits.
- **Autopatrolled**: Editors whose articles are automatically marked as "reviewed" by the system.
- **Administrators (or sysop)**: Editors who are granted by the community exclusive access to several tools, to allow them to carry out certain functions.
- **Bots**: Virtual users that run automated tasks on pages. Usually, they are programmed by other users to perform simple and repetitive tasks, much faster than a person could.

Users have a "User page" where they can give basic information about themselves and their work on the platform and related projects. Users can also specify some information about themselves on the preference page, such as their username, their signature, or their gender.

A feature of relevant importance in this research is UserBoxes.[1] They are graphical decoration users can add to their page and consists of a colored rectangle with some text and, sometimes, a picture, used to provide a notice about a user. UserBoxes are often embedded into a user page through a pre-made template, making them similar between multiple users.

2.1.2 Wikipedia talk pages

Talk pages can be of two kinds: "Article talk pages" and "User talk pages".

"Article talk pages" is the second biggest namespace in Wikipedia, after "Articles". Each Wikipedia article has an article talk page, which is an administrative page where editors can discuss improvements to other articles or Wikipedia pages.

Any user page has a corresponding "user talk page" that can be used by other users to send messages or discuss with other users.

The focus of this research is on these pages since users can express their emotions more freely and interact with each other directly. Talk pages are divided into threads, which groups user's posts on a topic of discussion. Each post is made by a user, and it can start a new thread or reply to an existing one. At the end of a post, a user is highly encouraged to leave his username and the current date. Several guidelines on these pages structure can be followed by editors, to have consistency on all pages.

Figure 2.1: Example of a discussion on a talk page

2.2 WikiConv Dataset

WikiConv is a multilingual corpus encompassing the history of conversations on Wikipedia Talk Pages, including deletion, modification, and restoration of comments, called actions.[2] The version we used includes all conversations extracted from the 2020-01-04 Wikipedia dumps of English, Spanish, Italian, and Catalan.

Each language version of the dataset is split into smaller files compressed through gzip. The files contain an action for each line.

Language	Number of actions	Number of files	Size of files	Total size
English	235 Millions	50	2.5 GB	485 GB
Spanish	20 Millions	5	2 GB	41 GB
Italian	17 Millions	5	2 GB	41 GB
Catalan	2 Millions	1	2 GB	7 GB

Table 2.1: Size of the WikiConv Dataset in various languages.

An edit of a page by a user is called revision, and can be represented by several action of four different kinds:

[1] `https://en.wikipedia.org/wiki/Wikipedia:Userboxes`
[2] `https://github.com/conversationai/wikidetox/tree/main/wikiconv`

- **Creation**: An edit that creates a new section in a Wikipedia page.
- **Addition**: An edit that adds a new post to a thread.
- **Modification**: An edit that modifies an existing post.
- **Deletion**: An edit that removes a post.
- **Restoration**: An edit that restores a previously removed post.

Each action is a JSON object that contains several pieces of information, as listed on the dataset documentation. The most relevant data for this study is: the page title and page id; the user name and id; the text of the post underlying the action; the timestamp of the revision and the type of action.


```json
{
    "id": "652113713.0.0",
    "revId": "652113713",
    "type": "ADDITION",
    "conversationId": "652113713.0.0",
    "pageTitle": "Talk:Industrial Policy Frame work for State of Telangana 2014",
    "content": "this is a very important thing.\n",
    "cleanedContent": "this is a very important thing.",
    "user": {
        "id": "[redacted]",
        "text": "[redacted]"
    },
    "timestamp": "2015-03-19T19:08:18Z",
    "pageId": "45718490",
    "ancestorId": "652113713.0.0",
    "authorList": [
        {
            "id": "[redacted]",
            "text": "[redacted]"
        }
    ],
    "comment": "i changed it a bit",
    "score": {
        "toxicity": 0.008058453910052776,
        "severeToxicity": 0.0010886316886171699,
        "profanity": 0.004147269763052463 5,
        "threat": 0.003460822394117713,
        "insult": 0.003118146210908889 8,
        "identityAttack": 0.004319264553487301
    },
    "pageNamespace": 1
}
```


Figure 2.2: Example of an action contained in the English WikiCovn

2.3 Emotional analysis

Emotional analysis tries to understand which emotions a user expresses while writing. Thanks to the WikiConv dataset, we can analyze what users wrote on Wikipedia talk pages and try to figure out their emotions through emotional analysis.

The spectrum of human emotions can be categorized in several ways. We choose the classification introduced in "Classifying emotion: a developmental account" [7], which suggests four types of basic emotions: fear, anger, joy, and sadness. Another classification we used is the separation of emotions into two sentiments: positive and negative.

Thanks to premade emotional dictionaries, we can associate each word in a post to a set of emotions and sentiments. We took into consideration the results obtained by EmoLex and LIWC.

2.3.1 EmoLex

EmoLex, or NRC Word-Emotion Association Lexicon [5], is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive).[3] The annotations were manually done by crowd-sourcing,

[3] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.html

and it is available for free for academic purposes. It contains 14,182 words, and its contents have been translated in over 100 languages with Google Translate since it has been shown that a majority of effective norms are stable across languages. There are present the four languages we are interested in: English, Spanish, Italian, and Catalan.

2.3.2 LIWC

Linguistic Inquiry and Word Count (LIWC) associate words to psychologically relevant categories and is considered by many the "gold standard" of in computerize text analysis [6].[4] To use LIWC an academic license is needed.

The latest LIWC dictionary was released in 2015 and consists of 6,400 words, linked to a set of relevant categories. LIWC is thought to be used through a processing module that takes as input a text and outputs percentages of different relevant categories it has found, but a dictionary file like the one offered by EmoLex can be extracted and used.

LIWC is available in different languages, including English, Spanish, and Italian. Currently, there is no version for the Catalan Language.

We got access to LIWC only in the last days of our research, and the results are not as thorough as the ones we got from EmoLex and are only used as validation.

[4] `http://liwc.wpengine.com/`

3 Methods

This project was part of a team effort. My contribution has been greatly influenced by other developers in our team, and the information we got were shared among us. For this reason, several aspects were a work of more than one person. In this chapter I will focus on what I personally did, but I will have to refer to work done in collaboration with others.

To share our projects and our implementations we used a GitHub organization where all our code can be found and used by others.[1]

3.1 Categorizing users

Analyzing and making public data about a single user can be considered a violation of their privacy. We decided to group users into categories identified by their gender and role in Wikipedia and to publish only results relative to groups and not individuals.

Obtaining this information is not straightforward.

3.1.1 Users' gender

In Wikipedia, there are two main ways with which users can specify their genders: Userboxes and profile settings.

UserBoxes

A UserBox, or UBX, is a graphical component that can be added to a user page. Some users have decided to express their gender through a corresponding UBX. There exist several genders related UBXs, grouped into 4 categories: "masculine", "feminine", "non-binary" and "others".[2]

Thanks to Wikipedia APIs, we can list all pages that contain those templates, collecting a list of user pages, and thus usernames, that decided to express their gender as stated in the corresponding UBX.

As we will see in section 4.1.1 this technique retrieved a relatively small amount of data if compared to profile settings.

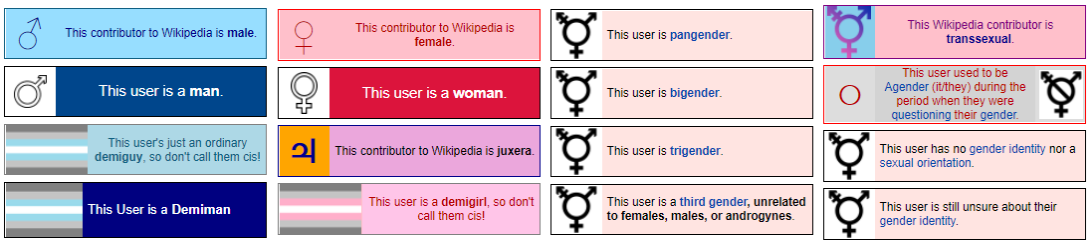

Figure 3.1: Examples of different UserBoxes used by the community

Profile Settings

Each Wikipedia user has access to a settings page, where some preferences and information can be specified, including a user's gender. The user choice is restricted to "Male", "Female" and "Unknown", where "Unknown" is the choice selected by default when someone creates an account.

[1] https://github.com/WikiCommunityHealth

[2] https://en.wikipedia.org/wiki/Wikipedia:Userboxes/Life/Gender

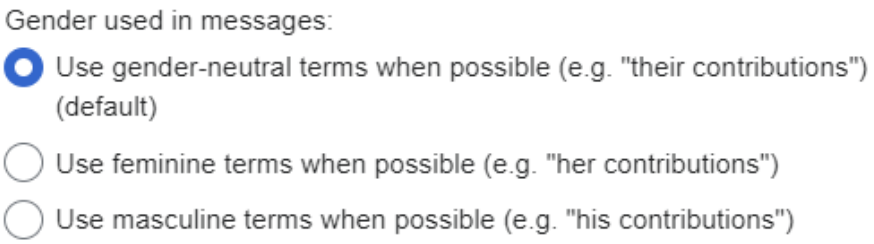

Figure 3.2: Settings page where users can choose their gender

Wikipedia APIs returns a list of information about a user, among which their specified gender in the profile settings.[3] We collected this and other information for all users for each language of Wikipedia.

Users' Role

Users have been categorized based on their user groups. We considered the most common groups, which are Admins (or sysop), autopatrolled, and registered users.[4] This information has been retrieved from a dataset generated by another team member where all information about a user are collected, including their roles. Thanks to this dataset we could identify some users as a bot and remove them from our analysis.

3.2 Lookup tables

The WikiConv dataset only provides information about actions. For our analysis, it was useful to have access to information about specific users and pages as a reference.

We decided to implement two lookup tables with basic information about each user and each page in a MongoDB instance,[5] a No-SQL database that stores data in JSON documents inside the collection. We choose this database because it allowed us to store and easily query our data without changing its format. To generate them, all the WikiConv dataset has been loaded into the database, then, a query, aggregated all information by the user and by page, and computed the new information seen in Figure 3.3.

[3] `https://www.mediawiki.org/wiki/API:Users`
[4] `https://en.wikipedia.org/wiki/Wikipedia:User_access_levels`
[5] `https://www.mongodb.com/`

```
{
    "_id": "100000",
    "totalActions": 412,
    "pageTitle": "Talk:English Civil War/Archive 1",
    "nrOfRevisions": 127,
    "typeOfActions": { "ADDITION": 90, "MODIFICATION": 53, "DELETION": 161, "RESTORATION": 88 },
    "scoreActions": {
        "toxicityCounter": 11,
        "severeToxicityCounter": 0,
        "profanityCounter": 7,
        "threatCounter": 0,
        "insultCounter": 3,
        "identityAttackCounter": 0,
        "toxicityCounterRatio": 0.02669902912621359,
        "severeToxicityCounterRatio": 0,
        "profanityCounterRatio": 0.01699029126213592,
        "threatCounterRatio": 0,
        "insultCounterRatio": 0.007281553398058253,
        "identityAttackCounterRatio": 0
    },
    "activityDate": {
        "firstEdit": { "$date": "2002-10-06T09:53:35.000Z" },
        "lastEdit": { "$date": "2006-07-23T12:47:29.000Z" },
        "activeDays": 1386,
        "editsPerDay": 0.2972582972582973
    },
    "workedByUsersCount": 46
}
```


Figure 3.3: Examples of the document representing a Wikipedia page in the pages look-up table

3.3 Sorting the WikiConv dataset

The WikiConv dataset is composed of compressed files, each of them is about 2GB in size. Each file contains a random subset of actions of the entire dataset, and it is not possible to know in which file a specific action is contained.

We chose to sort the dataset in four different ways. Each sort is defined by a sorting feature. For each sorting feature, a set of intervals had to be chosen such that:

- Each possible feature is part of one and only one interval.
- All intervals have a similar number of actions associated with them.
- All intervals are sorted such that if an interval is greater than another, then all its features are greater than the features of the other interval.

All the process was done by decompressing the file "on-the-fly", without the need to store a plain version at any step of the process. This greatly increased performance, while reducing the storage space required for the process. The run time of the algorithm can be expressed as $O(N * M * log(M))$, where N is the number of intervals, M is the number of actions associated with an interval.

Of crucial importance was the choice of the dimensions of the intervals. An interval too big would result in a bigger M, a bigger file to sort, and thus a longer run time; whether a smaller interval would create a huge number of files to sort and a hardware bottleneck on the reading and writing performance, both during the sorting and during the analysis.

The four different sorting features are Users, Pages, Timestamp, and Reply-To.

Users Each action can be associated with the user that made it, which can be identified by its numeric id. If an action was made by an anonymous user, no user id is saved on the dataset, but we can use its IP address to identify it. This method is not guaranteed to univocally identify a user, since many users can share the same IP address, or the same user can change IP address over time. A small set of actions had no user id and no IP address.

They were associated with a null user for the sake of the algorithm but were not analyzed after.

Page To sort by page, we used the page id saved on each action as a sorting feature. The intervals are ranges of dimension 2 million, which covers all possible pages' ids.

Timestamp To sort by timestamp, we used the timestamp saved on each action. Actions from the same revision have the same timestamp, and we used their id as a secondary sorting feature since ids are ordered inside a revision. The intervals range from the start to the end of each month analyzed by the dataset.

Reply To The reply to is a field that is not present in the original dataset. It was computed by another team member, by reconstructing the discussion tree of each page and assigning as "reply to", the id of the user that the action was addressed to. This field is useful to understand which emotions are addressed to which user.

The intervals we used are the same as the ones we used for the user id.

Language	Lines	Size	Compressed size	Time to sort
English	235 Millions	485 GB	129 GB	12 h
Spanish	20 Millions	41 GB	11 GB	1:30 h
Italian	17 Millions	41 GB	11 GB	1:30 h
Catalan	2 Millions	7 GB	2 GB	0:30 h

Table 3.1: The table shows the number of actions in each dataset, the size of the compressed and uncompressed dataset and how much it took to compress.

3.4 Minifying the WikiConv dataset

The WikiConv contains a lot of information that is not used during our analysis. We choose to generate a new parsed dataset containing only the information we needed. For each action, we kept id, action type, page title and id, username and id, and timestamp. The text content was parsed into an array of integers, where each integer is the number of words in the text associated with an emotion according to the lexicon dictionary.

Analyzing these versions of the dataset was much faster since the words were already associated to an emotion, and thanks to the reduced size of the file. It was also simpler to move and store the files.


```
{
    "id": "703157392.2774.2708",
    "type": "ADDITION",
    "pageTitle": "User talk:[redacted]",
    "pageId": "[redacted]",
    "user": {
        "id": "[redacted]",
        "text": "[redacted]"
    },
    "timestamp": "2016-02-03T21:16:45Z",
    "pageNamespace": 3,
    "emotions": "82,13,3,0,1,0,1,2,0,1,5,"
}
```


Figure 3.4: Examples of the document representing an action in a minified dataset

Language	Sorted dataset size	Minified dataset size
English	81 GB	3.2GB
Spanish	6.8 GB	0.32 GB
Italian	6.1 GB	0.30 GB
Catalan	1.4GB	0.04 GB

Table 3.2: The table shows the difference in size from the sorted dataset to the minified dataset.

3.5 Analyzing the WikiConv dataset

All the analysis tools have been written in Python 3.[6] Python is an interpreted programming language suited for data science. It has a plethora of useful libraries for the tasks we are interested in and offers great flexibility.

We developed a python module that offers a command-line interface to allow a user to analyze either the sorted WikiConv dataset and the minified sorted WikiConv dataset. Through dependencies injection, the module can load an "analyzer": a service that offers some APIs for the module to compute a different kind of analysis.

The module computes several tasks. Firstly, it reads a specified set of files, decompressing them on the fly, if necessary. Then it goes through all actions in the file, grouping them by their sorting feature. Since the sorting feature is sorted, once a different feature is read, the previous group is completed and can be passed as input to the analyzer which does its specific computations and remove from memory. The module also handles parallelization running each file in a different thread. The user can specify if he wants to use parallelization and the maximum number of parallel threads running.

An analyzer handles every possible analysis. They can be developed independently but need to implement a set of APIs called by the module. They are used to pass the actions groups and to synchronize the progress among different threads.

We developed several analyzers to handle different tasks, the important ones are described here:

Minifier This analyzer allowed us to reduce the dataset structure as described in section 3.4. It took each line, calculated its minified version, and generated the new file version.

Reply To It was used to create a new feature in the dataset called "replyTo" described in section 3.3. This feature identifies which users are an action addressed to.

Reply To Mean and Var Analyzer: It analyzed the mean and variance of emotions in each activity group. The output is a JSON file containing the total mean and variance of each emotion contained in the analyzed text and the number of actions analyzed. The results were also calculated and saved for each month contained in the action group. This gave us global information about a user or a page through time. It was also possible to use this analyzer to compute the same information for the whole dataset, giving us a reference point and a view of the whole Wikipedia in a particular month.

Word Cloud Word Cloud: This analyzer created a graphical representation of the most used words by a user, by a page, in a month, or Wikipedia as a whole. Seeing what words, the lexicon analyzed and associated with different emotions was difficult to verify manually, and having a graphical representation of its work was useful. It was used to verify and validate our works and the dataset we used and to illustrate our works to others simply and effectively.

Emotions to DB This analyzer reads all emotions and generates the metrics to add to a Postgres database. We created metrics about pages and users' groups for any month in the dataset. These metrics were saved in a SQL database and were then published.

[6] `https://www.python.org/`

Emotions to JSON This analyzer, similarly to "Emotion to DB", generates metrics about users' groups and pages, but saves them in a JSON file. This is mostly used to rapidly share information among team members and is used as input for other scripts, such as the one that draws the charts.

Emotions to CSV Similarly, to "Emotions to DB" and "Emotions to JSON", this analyzer generated metrics about users' groups and pages in CSV. These files were used by a researcher to develop a project of data visualization. We choose to use CSV since it is a simple and straightforward file format, that can be used without any special technical abilities.

3.6 Metrics

Metrics are the core result of this study. They can be used to characterize users and pages by their emotions and can be exported and incorporated into other projects. It is important to make all metrics useful on their own and should be compared with each other. Each team member had different metrics and different approaches to compute them, but we tried to make them as similar as possible following this guideline:

- Each metrics value should refer to a users' group or a page
- Each metric value should refer to a month
- For each metric value there should be three linked metrics:
  - **Normalized**: the value normalized with a set of other metrics
  - **Accumulated**: the current month metric value summed to all the same metrics of the previous months.
  - **Accumulate Normalized**: the accumulated value normalized in the same way the normalize value was calculated.

To facilitate this process a Python package was developed, with the help of a team member. It helps our team members to format the metrics to follow our guidelines and push them into a database. The package can be installed via pip in any Python 3 project.

4 Results and Discussions

4.1 Users' gender

We retrieved information about a user gender in two ways: UsersBoxes and User's settings. These data were then used to categories users' in groups.

4.1.1 UserBoxes

Analyzing UserBoxes we retrieve relatively few data. Most Wikipedia users do not spend much time on their user page. To insert a UserBox it is required an effort by the user, that most are not willing to take. Gender UserBoxes allow a wide personalization and more can be added as needed. All genders are macro-categorized as Masculine, Feminine, Non-Binary and Other.

Genders categories	**Number of users that used one of those UserBox**
Masculine	10,489 Users
Feminine	2,045 Users
Non-Binary	347 Users
Other	218 Users
Total	13,099 Users

Table 4.1: The table shows the number of users that implemented a UserBox of a particular category.

This approach may be subjected to some biases. Users that do not feel included in the limited gender selection of the Wikipedia user profile, may tend to express their identity through this feature, while users correctly represented on their settings, may not feel the need to use this feature. Another known bias, particularly on internet community, is the tendency, from female users, to not specify their gender or specify it as male. This is due to the fact that female users are subjected to discrimination in these platforms and tend to hide their true gender.

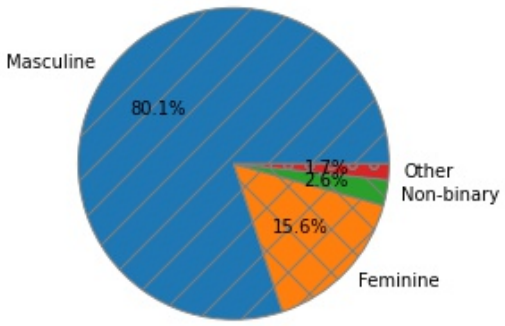

Figure 4.1: Percentage of editors with a UserBox for each gender group

We choose to not use this data, because it was statistically too small if compared to the profile

settings and would generate groups to with few to no data at all.

4.1.2 Profile settings

Profile settings is the most used feature to express a user gender. Its options are limited to male, female and unknown, where the unknown option is set by default. Most users never change this option, and are thus categorized as unknown, so, this specific label does not represent any valuable information.

This information is subjected to the same biases stated above, but it was the best we could do with the resources we had. Our results show a similar distribution of female and man to more accurate studies such as "Gender differences in Wikipedia editing" [1]. We can confidently say our results are a good representation of male and female users in Wikipedia.

We only needed data about English, Spanish, Italian and Catalan Wikipedias, but we decided to collect data for all Wikipedias. The results for the most common Wikipedia languages are shown in table 4.2.

Wikipedia	**Female Users**	**Male Users**	**All Users**
English	122,445	604,633	40,526,936
Spanish	30,774	106,269	6,071,869
German	14,371	79,010	3,603,382
French	9,275	48,531	3,968,600
Russian	55,065	173,019	2,891,204
Italian	7,731	39,037	2,060,494
Catalan	1,499	5,145	373,672

Table 4.2: The table shows the number female, male and total user for the most used Wikipedia.

This data was used in to generate two user groups: female and male.

We analyzed the gender distribution for users with different edits counts, since it is a simple way to identify a user level of activities, and it can be said that users with more edits are more active. The results are shown in Figure 4.2.

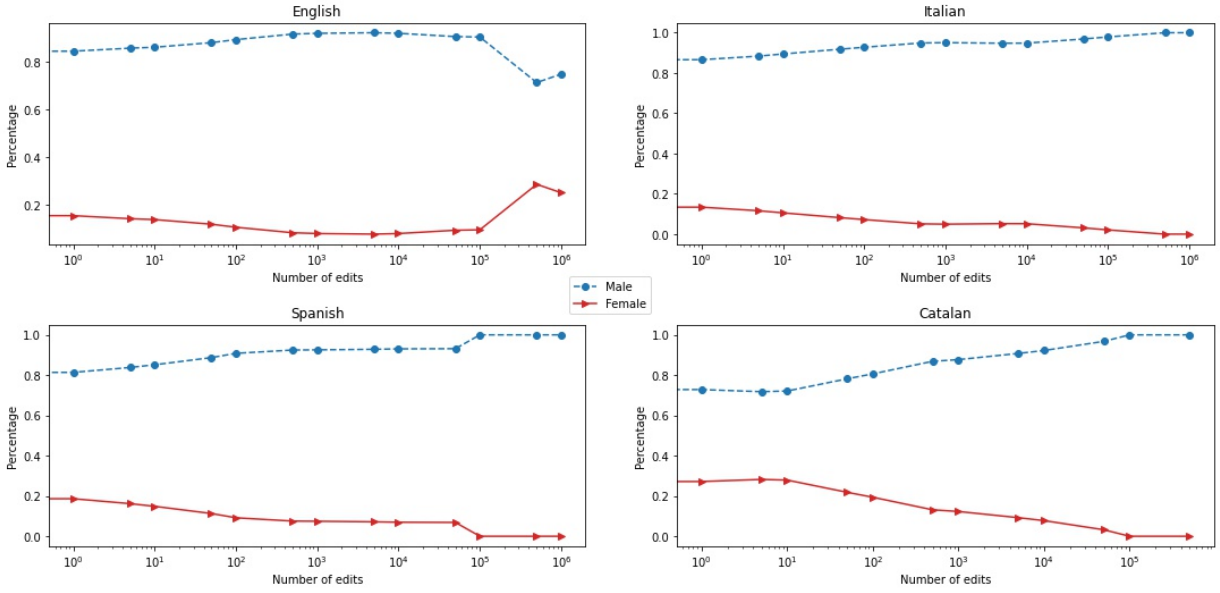

Figure 4.2: Percentage of male and female based on edit count, in four different Wikipedia. The users edit count is calculated based on a series of threshold, that a user can reach. The thresholds are: $0, 1, 5, 10, 50, 10^2, 5 \cdot 10^2, 10^3, 5 \cdot 10^3, 10^4, 5 \cdot 10^4, 10^5, 5 \cdot 10^5, 10^6$

Our analysis compares male users and female users to a reference made by all users. All data have been shared in our group, but it was not made publicly available to avoid users' privacy violations.

As seen in Figure 4.2 male editors are the vast majority of users, and the difference between male and female editors tend to grow with the edit count.

4.2 Emotions

Emotions are the core of this study. We analyzed the emotion expressed by users on the article and user talk pages. The analysis was done with EmoLex.

WikiConv	**Action analyzed**	**Words analyzed**
English	4,163,088	1,178,703,180
Spanish	469,051	95,987,997
Italian	245,886	45,248,302
Catalan	28,557	7,264,086

Table 4.3: The table shows the number of actions and words analyzed foreach WikiConv Dataset during our analysis.

Our first result was to calculate how much each emotions was expressed in the language we analyzed. We estimated the percentage of words which express an emotion or a sentiment over all words analyzed that were contained in the emotion dictionary. The results are shown in Figure 4.3.

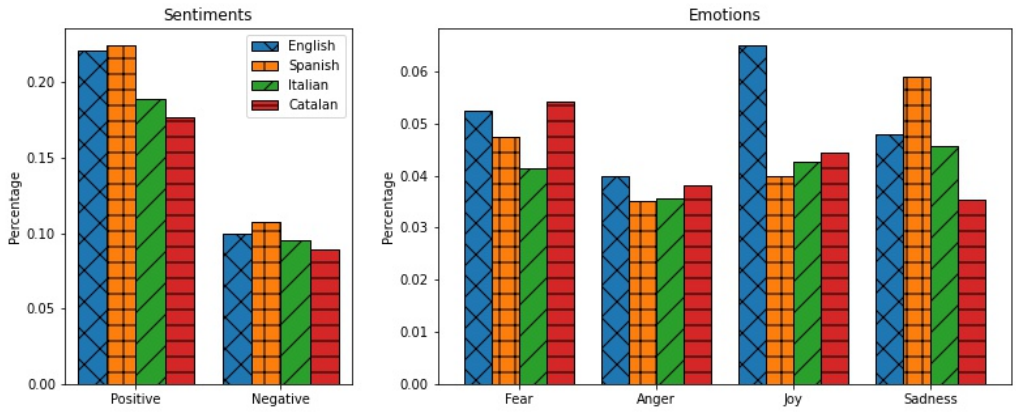

Figure 4.3: The tables shows the percentage of emotions and sentiment expressed in all talk pages on different Wikipedia

4.2.1 Word Clouds

We analyzed a great quantity of data. It was necessary to have a visual representation of what we were doing. We choose to represents the most common words identified as emotions by our dictionaries in a "Word Cloud".[1] Some sample results can be seen in Figure

[1] `https://amueller.github.io/word_cloud/`

Figure 4.4: This are two example of "Word Clouds" extracted from the Italian WikiConv of November 2019. The dimensions of a word in these images is directly related to their presence in the dataset. The number on the title is the total number of analyzed words that belong to that particular emotions.

4.2.2 Users

We can use users' group previously defined to analyze their different emotional response. For each user we calculate the average value for the emotions over all its actions and averaged this with all other users in a group. The results can be seen in Figure 4.5 and Figure 4.6.

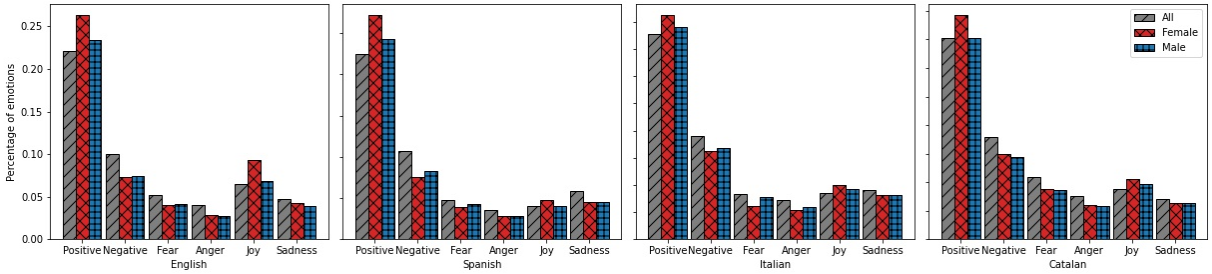

Figure 4.5: The tables shows the percentage of emotions and sentiment expressed by female, male or all users

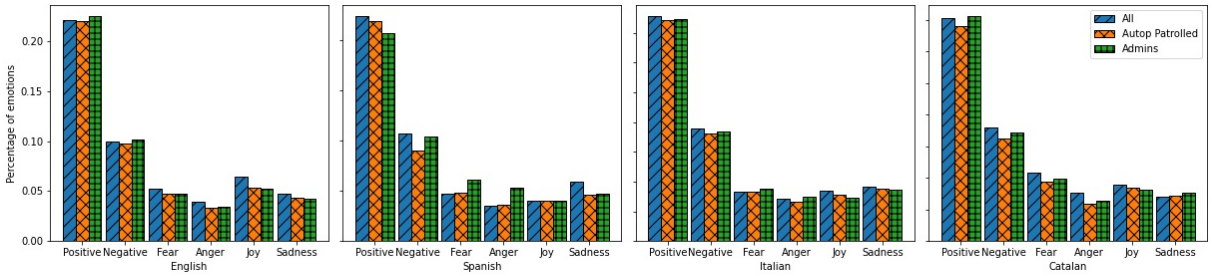

Figure 4.6: The tables shows the percentage of emotions and sentiment expressed by admins, auto patrolled and for all user

In figure 4.5 we can see that female editor have a tendency to express more positive sentiment than male. This is was also show in previous studies [3] [2].

We are interested in users life-cycles, from when they join Wikipedia to the moment they leave. To better understand how users feels during their life as a Wikipedia editor we analyzed all emotions they express. We took an average of a user action in each month after their complete their first action, and averaged them with all users in a group or with all Wikipedia's users. We took into consideration that the number of active users decrees over time. We also analyzed the emotions a user received from others.

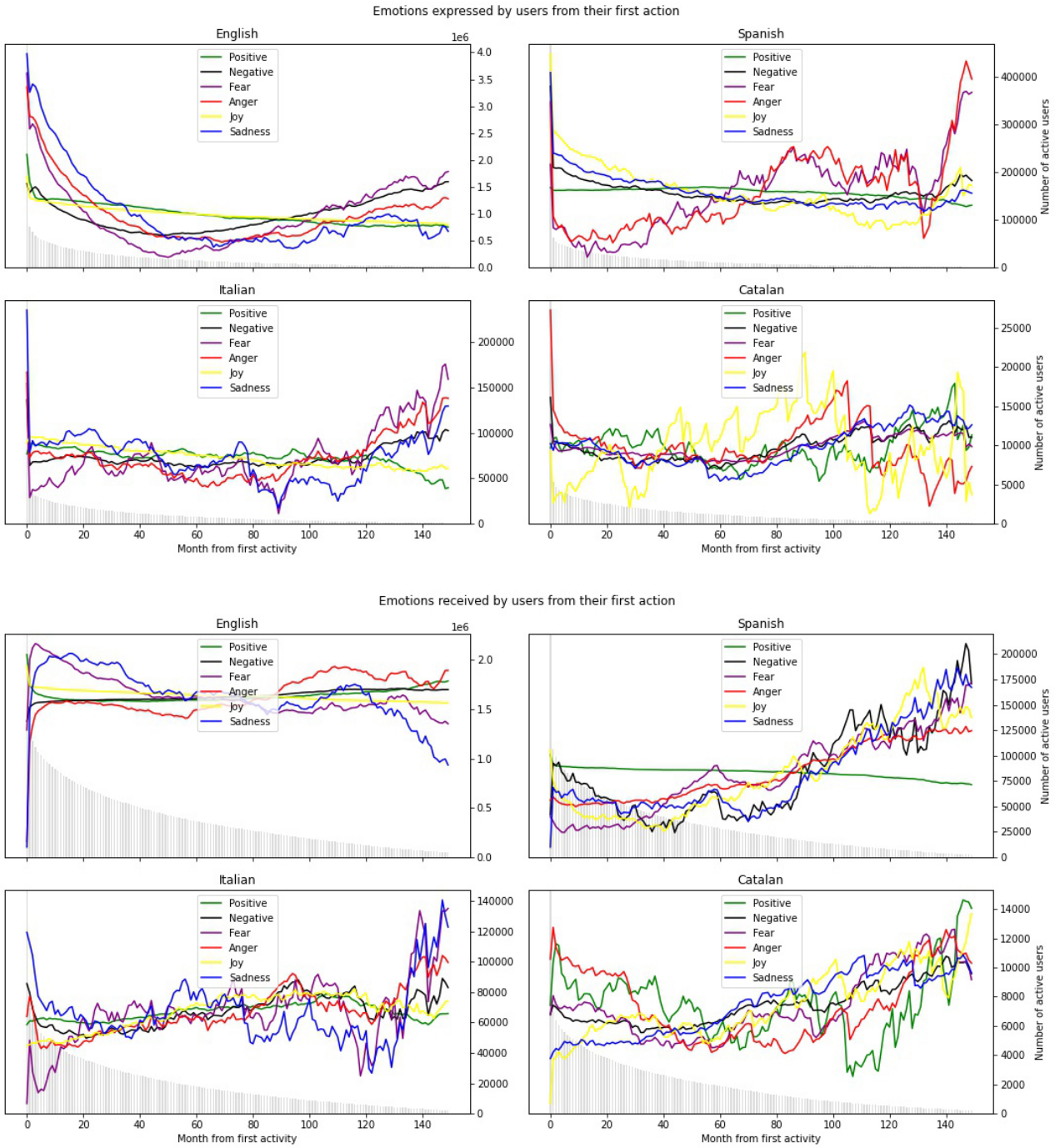

Figure 4.7: The two tables shows the normalized value of each emotion and sentiment analyzed for four languages. The bars in the background represent the number of active users it and the x axis are the month from a user first activity. The first table shows the emotions expressed by a users in they actions and the second the emotions received through reply to their posts.

Life-cycles can also be analyzed for different users' groups and compared between them. This can be useful to understand differences between different users.

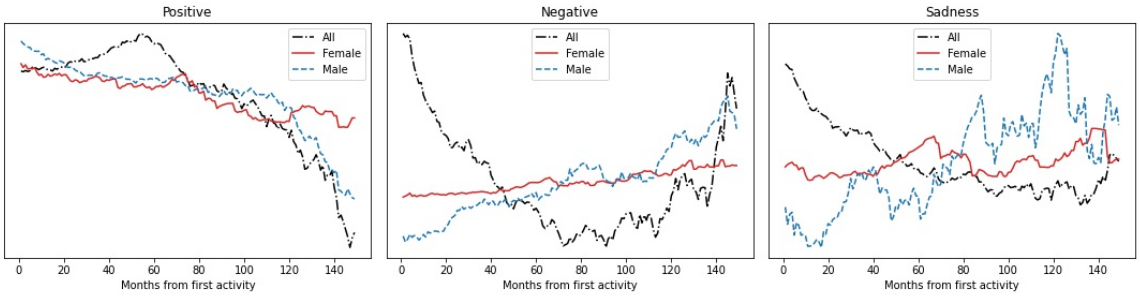

Figure 4.8: The two tables shows the normalized value of different positive, negative and sad words used by users of different gender from their first action.

To understand how external events influence users on Wikipedia it is useful to see variation of emotions over time. For this reason we analyzed user emotions for over each month from the opening of Wikipedia, in a similar way we analyzed user life-cycle. It is important to remember that, in the early days Wikipedia was less used, and we have less data. Wikipedia reached a good level of popularity around 2007.

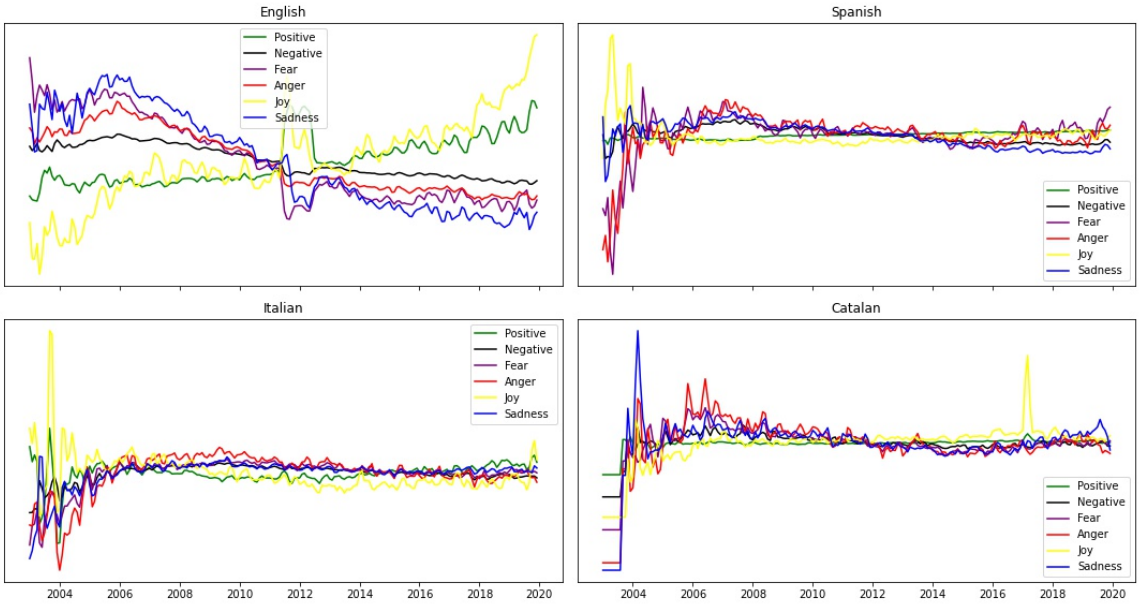

Figure 4.9: The table shows the normalized value of all analyzed emotions over time. The emotions are sampled each month

4.2.3 Pages

Pages were analyzed similarly to users. In this case there were less privacy concerns since pages aggregate contribution from different users. We did not made any groups of pages. Smaller pages do not have enough information to be considered statistically useful, but bigger pages could be analyzed. Since it did not increase significantly our calculation time, we decided to analyze all pages and save the metrics we generated to our database.

Every page has a metric for each emotion and sentiment from any month of its life.

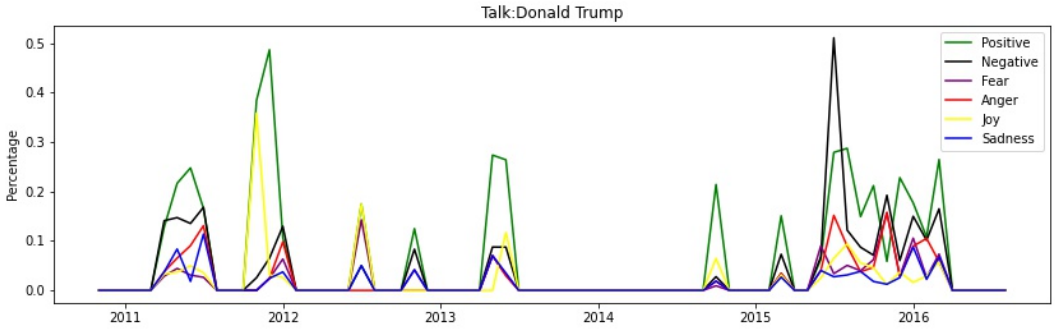

Figure 4.10: The table shows the emotions expressed by editors on the talk page of the article related to Donald J. Trump

5 Infrastructure

5.1 Storing the dataset

We stored the dataset in multiple ways, each better suited for different tasks.

The original WikiConv, its four sorted copies, and the relative minified versions were stored as compressed gzip files on a hard disk and backed up in Google Drive. This hard drive was accessible from all the computers we used to perform the analysis.

A copy of the original dataset was decompressed and loaded in a MongoDB instance, where the look-up table was also created. MongoDB is a non-relational database that uses a JSON structure for its data, making it a good fit for the WikiConv data structure. It was used to launch simple queries on the dataset and retrieve immediate results. MongoDB was also useful for other related projects, which used our data in their analysis, and could easily use it.

5.2 Storing our results

Results are stored in four ways: A Postgres database,[1] A mongoDB instance, JSONs files and CSVs files.

Among all members of our group, we choose to use a similar format for all our metrics. All the relevant metrics we generated can be saved in a Postgres database and do not contain sensitive information. Each Table collects information about a single language analyzed. The rows every month are analyzed for each page and user group.

MongoDB was mostly used for the look-up tables. It was also used to extract the information needed during our analysis that was made by other team members and was stored on a MongoDB instance.

JSON files were used to save simple results, that could be shared or used to plot charts.

CSV files were mostly used to save data used in our analysis that did not need to be loaded to a proper database and needed to be read quickly from a file. This is also the format the emotion EmoLex used.

5.3 Public API

We want to make publicly available the metrics we calculated, but the dataset is large and, usually, only a small portion of it is necessary for any related work. Forcing people to load the whole dataset can be time-consuming and take up lots of resources, setting a high bar for anybody interested in using our data. To reduce this stress, we decided to make a public API endpoint. People can now query our dataset without the need to install it on their machine and retrieve only the information they need.

To maximize the flexibility of our APIs results, we implemented GraphQL, a query language made for APIs.[2] It allows users to write and send to our backend their query. We can validate and execute their request and return the data in the format they asked for. GraphQL also offers a landing page, reachable from any browser where user can compose their query and are helped by the UI to understand our database structure.

To link GraphQL to our Postgres database we used Hasura, a software that automatically maps a Postgres database to GrapQL. This solution allowed us to use GraphQL, without the need to implement a custom backend. We configured Hasura to allow anybody to read our data, but without the possibility to alter it.[3]

[1] `https://www.postgresql.org/`
[2] `https://graphql.org/`
[3] `https://hasura.io/`

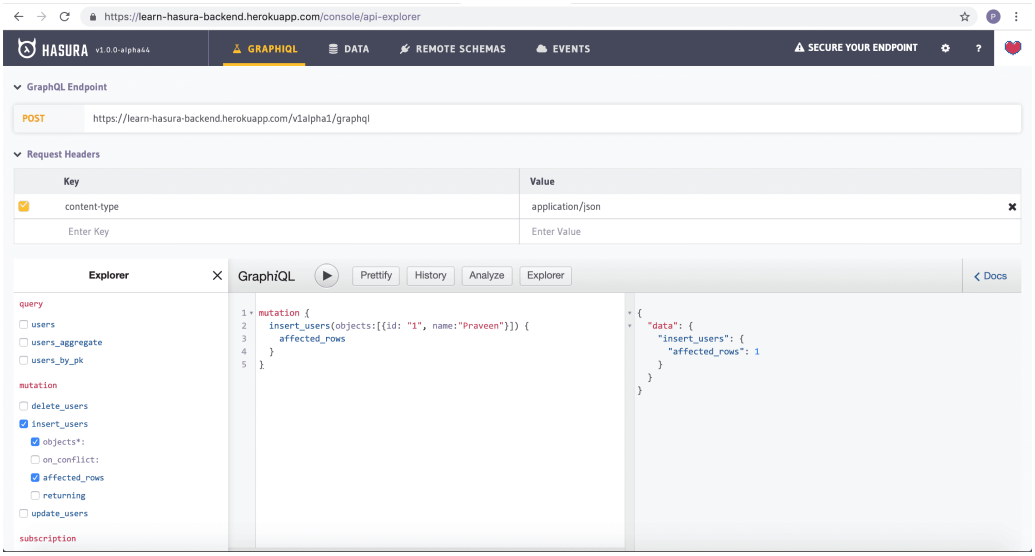

Figure 5.1: Hasura landing page, where a user can query our database

5.4 Web Application

We developed a web application to show chart combining multiple metrics. This application was developed in TypeScript[4] using the Vue.js framework.[5] For the charts we used the JavaScript version of Plotly.[6]

The Web application provides a simple and accessible tool for anybody that wants to read and plot out metrics. It is accessible from a browser, and through a graphical user interface allows users to select which metrics to plot and compare. Thanks to the metrics being saved similarly, results from multiple projects from our team can be implemented.

The web application was developed in Typescript using Vue.js 3, a progressive JavaScript Framework. All code developed in Vue.js runs on the user browser, avoiding the need for us to maintain a complex infrastructure active, it is only required to serve the static HTML and JavaScript files from the server. To retrieve the data needed by a user, the web application queries the Hasura instance. This is done through Apollo,[7] a GraphQL implementation for Vue.js.

The charts were made with Plotly JavaScript, a charting library, which allows us to build complex charts with a high-level implementation. Charts show different metrics simultaneously and can be personalized by users.

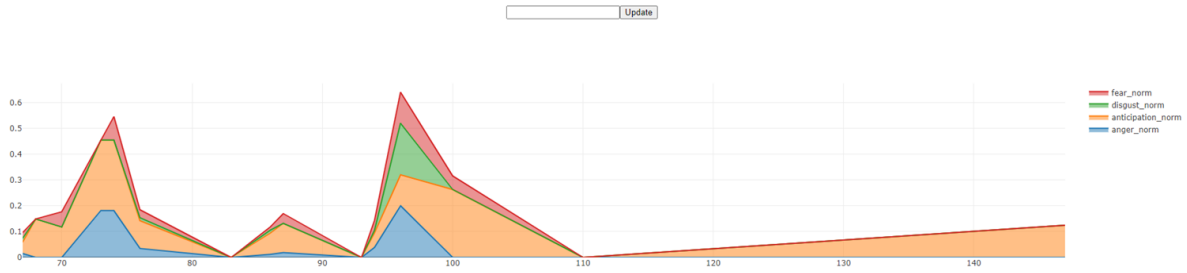

Figure 5.2: A page of the web app where a chart is showing severla metrics chosed by a user

[4] https://www.typescriptlang.org/
[5] https://vuejs.org/
[6] https://plotly.com/javascript/
[7] https://www.apollographql.com/

5.5 Deploy

To deploy our solution, we decided to use Docker.[8] It is a software that uses virtualization technology to deliver software in packages called container. Containers are isolated from one another and bundle their own software. Thanks to this technology we can deploy our database, Hasura, and our web application in three different containers. We can deploy our solution without worries about dependencies and compatibility, moreover, all our software is isolated and can be updated with continuous integration.

The Hasura container is public and anybody can implement their application using our data.

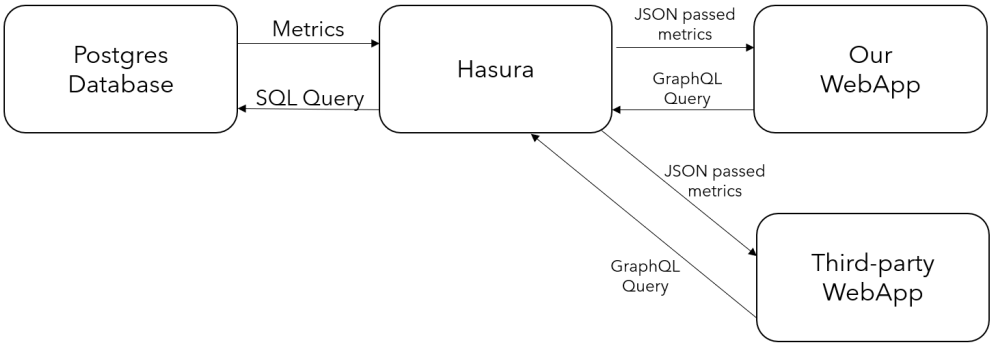

Figure 5.3: This is a graphical rappresentations of out docker containers and a possible third-party application with their communication

[8] `https://www.docker.com/`

6 Conclusions

With this project we have an overview of emotions expressed in Wikipedia discussions, but much more can be done. With the help of other team members, we plan to join our metrics and generate a better representation of an editor's life cycle. Future work on emotions can also be done. We want to implement other emotion lexicon, such as LIWC, to have a comparison with other system, and use information from other team members to better define users' group.

The current dashboard is a bare-bone version of what we have in mind. We want to keep improving it, joining metrics from different projects and giving users greater freedom on the choice of data thy want to visualize.

All the information we gather will soon be outdated. For this reason we are implementing a system that automatically updates our metrics with the latest Wikipedia dump. We also plan to extend this analysis to other languages.

All the code wrote, is public and was made to be modular, giving the possibility to others to continue this work.

Bibliography

[1] Judd Antin, Raymond Yee, Coye Cheshire, and Oded Nov. Gender differences in wikipedia editing. In *Proceedings of the 7th international symposium on wikis and open collaboration*, pages 11–14, 2011.

[2] Daniela Iosub, David Laniado, Carlos Castillo, Mayo Fuster Morell, and Andreas Kaltenbrunner. Emotions under discussion: Gender, status and communication in online collaboration. *PloS one*, 9(8):e104880, 2014.

[3] David Laniado, Andreas Kaltenbrunner, Carlos Castillo, and Mayo Fuster Morell. Emotions and dialogue in a peer-production community: the case of wikipedia. In *proceedings of the eighth annual international symposium on wikis and open collaboration*, pages 1–10, 2012.

[4] Marc Miquel-Ribé, Cristian Consonni, and David Laniado. Wikipedia editor drop-off.

[5] Saif M Mohammad and Peter D Turney. Nrc emotion lexicon. *National Research Council, Canada*, 2, 2013.

[6] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[7] Alexandra Zinck and Albert Newen. Classifying emotion: a developmental account. *Synthese*, 161(1):1–25, 2008.