Universiteit Leiden
Opleiding Informatica

To what extent can we use Machine Learning to predict the emotion of a person after reading a Dutch news article?

Name: Kirandeep Kaur
Date: 24/06/2020
1st supervisor: Peter van der Putten
2nd supervisor: Jasper Schelling

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

In this thesis, we aim to predict the emotions of human beings after reading Dutch news articles. The news articles have been pre-processed and come from various news channels. The chosen volunteers will label the articles with the emotion felt after reading, using the Geneva Emotion Wheel method. Other emotion labeling methods will also be discussed. On the resulting data, machine learning techniques will be applied to develop models that predict the most likely emotion that an article could provoke. The process involved in developing these machine learning models will further be discussed and the models will be compared with each other, to finally come to a conclusion to what extent the emotion felt by a person can be predicted. This research is part of a bigger project and is intended to contribute to the main goal of that same project, Bursting the Bubble. This project aims to decrease polarization and increase diversity in the Netherlands.

Contents

1 Preface

This thesis marks the final work of my Bachelor's degree in Computer Science at Leiden University. It serves as the documentation of my research on human emotions and media bias in current times. During this study conducted over the span of 6 to 9 months, I got to learn a lot about Machine Learning and Data Mining. I got the opportunity to combine multiple fields such as Literature and Psychology with Computer Science. Managing a full time job while writing this dissertation was all about balancing work and still reading and delving into books and articles on the weekends, which eventually proved to be quite fun. Finding the right balance between my DevOps engineering job and the dissertation was all possible thanks to my supervisor Peter van der Putten, who gave me enough space and time to come up with new ideas and providing me with the right feedback so I could come up with my own.

2 Introduction

In the last few decades, media and news have formed an integral part of our daily lives. Politicians have become more and more aware of their public perception through media. Earlier, it was through news papers and radio broadcasts, but now there is access to the same news channels on laptops and smartphones. These news channels can be seen influencing our opinions and views on several topics. In fact, in a survey conducted by Gingerich in 2017 it was found that many individuals consider traditional paper news articles to be of higher quality and trustworthiness than other media formats such as social media [1].

As news media continues to strongly impact our views on various topics, understandably, many officials have tried to make use of this by using it as a means to make their image more attractive or positive in the eyes of the general public. Journalists are seen using selective information, events and words to publish news articles as they often let their personal views slip [18]. People with a certain opinion about a political party or celebrity are fed with news and information supporting those same views. This could lead to reinforcing personal biases or sinking into an ideological bubble.

The aforementioned ideological bubble is also known as a "filter bubble", a term coined by internet activist Eli Pariser [8]. In the book Filter Bubble, he states that the algorithms behind popular search engines such as Google create a unique universe of information for each of us, which fundamentally alters the way we encounter ideas and information. Big companies such as Facebook and Google feed you what they think you want. On the basis of your previous searches, a profile is created which you actually do not have access to. This created profile results in a reinforcement of beliefs, also called echo chambers.

In 2018, Panke[9] further elaborates on the reinforcement of personal biases. She states that "when individuals on both sides of an issue are polarised and only see their side of an argument, and continually see their opinion reinforced, a solution will never be reached because both sides will refuse to accept the opposing argument, and due to the deepening division between people and their opinions, individuals are struggling to agree to disagree" (p. 259). Continuing she also gives an example of liberals and conservatives in the United States. These liberals and conservatives often live in separate media worlds and show little overlap in the sources they trust for political news. Due to this separation, they are not exposed to the same broad information. This personalisation process could hinder a person from forming an objective and neutral opinion by not considering the content that they do not get to see.

However, Alex Bruns in 2019 argues that the influence of filter bubbles and echo chambers has been largely overstated and is actually a result from the general moral panic about the role of online and social media in society [10]. He explains how in the society that we live in, it would require a lot of effort to find only information that fits our existing worldview and claims the filter bubble and echo chambers to be a myth, "it is high time we cut through those myths and shifted our focus to the cognitive processes and ideological mindsets that produce such polarisation".

Understanding the importance of being able to look at topics from an objective point of view,

the bigger project my research is a part of: "Bursting the Bubble" focuses on how we could be exposed to a more information neutral environment by making use of Machine Learning and Natural Language Processing. This project wants to engage and challenge the public to different views and opinions, rather than reinforcing their prior ones. My research focuses on the sentiments that are associated with multiple Dutch news articles and can be used to further confirm or attest the existence of the aforementioned bubble that the society lives in.

As mentioned in the first paragraph, the news data can be easily accessed and collected on smartphones and laptops. For this research, the published articles of several news channels have been collected by a group of colleagues at the University of Leiden. Volunteers, chosen at random will then label the different news articles provided to them with the emotions they associate these news articles with. It will be assessed to what extent it is possible to predict how a person feels after reading an article. Does he or she feel elated, happy or angry? The news articles will be classified across the dimension; emotion. This research could prove to be useful in decreasing polarization and increasing diversity in the Bursting the Bubble project. If the emotion a person experiences can be predicted, it can be discussed what exactly makes them feel this way. If one can attain that information, it can be used to make the news articles that individuals are fed with more diverse and neutral.

It should also be noted, that the task of predicting emotions in people is an extremely complex one. Human beings express emotions in different ways including facial expressions, speech, gestures and written text. Moreover, the emotion felt by one person after reading an article does not have to be the same for another person reading that same article. Human beings experience many difficulties when trying to do this themselves, so for technology to accomplish this, even more obstacles are expected. This thesis presents the limitations, difficulties and risks associated with this technological approach.

This thesis is organised as following: section 3 explains some of the terms used in the research as we delve deeper into a few studies that have been published on related topics. Section 4 describes the data collection process, which is further used in sections 5 and 6. In these sections we also describe the process and decisions made throughout the conducted experiments. In section 6 we analyze how this research can further be useful in a bigger context. In section 7 we explain some of the ideas that came up in our minds during the course of the thesis and how they could provide basis for future research. Then, last but not the least, we draw our conclusions in section 8.

3 Background

In this section we will cover the different researches that have been conducted on topics that involve news emotion classification or sentiment detection. We will also describe the different terms that are used throughout the research.

3.1 Emotion classification

Previous studies have identified two approaches to analysing emotions; a categorical approach and a dimensional one.

In a categorical approach, emotions are characterised on a set of basic emotions that are cross-culturally recognisable. They are referred to as discrete because they can easily be recognised by facial expressions or biological processes. Ekman in 1969 identified six basic emotions; anger, disgust, fear, happiness, sadness, and surprise. These emotions are cross cultural, signifying they are universally recognised and are expressed similarly in the Western as well as Eastern culture [12]. In a more recent study Jack states that fear and surprise share a common signal, the wide open eyes, and anger and disgust share the wrinkled nose. Therefore, these should be combined resulting in only four basic emotions; fear, anger, joy, and sadness [11].

In a dimensional approach, the emotions are characterised on a dimensional basis in groupings. They are presented as a point or a region within a two- dimensional or multi-dimensional space and are therefore not a subject of assignment to a single category, but to many variables. Russel in 1980 posit the existence of two fundamental dimensions of emotional space: valence and arousal. The vertical axis representing arousal, the horizontal axis valence and the centre representing a neutral valence (medium level of arousal) [13]. To use a dimensional approach in our research, a dimensional model is constructed where emotions are defined according to one or more emotions. An emotion lies in two or three dimensions. The dimensions used are power and valence with a Geneva Emotion Wheel (GEW, [17]), which is an emotion classification model that we will explain in 2.3.

3.2 Sentiment Analysis

Sentiment Analysis, also known as opinion mining, is an interpretation and classification of emotions within textual information using natural language processing and text analysis. Subjective information is extracted, identified and studied. Many different machine learning or text mining, techniques have been applied for sentiment analysis, predicting a sentiment, positive or negative.

This has often been used in business settings to understand the social sentiment their brand carries. A study conducted by Arora, Deepali, Li and Stephen analyzed the reviews of major phone brands given on social media platforms such as Twitter, Facebook and LinkedIn [14]. The data can be benefitting in the sense it could potentially provide businesses more insight into users' responses to their products and services, so they can make improvements and gain an edge over their competitors. In 2019 [3] Taj, Meghji and Sheikh investigated sentiments present in textual information. The paper used a Lexicon based approach on news articles from the BBC news dataset and it was found that sports and business had more positive articles,

while entertainment and tech had more negative articles. In 2014, another study carried out by Doddi, Haribhakta and Kulkarni [7] focused on filtering out negative articles, "The objective of this project is to provide a platform for serving good news and create a positive environment. This is achieved by finding the sentiments of the news articles and filtering out the negative articles. This would enable us to focus only on the good news which will help spread positivity around and would allow people to think positively." (p.1). Aforementioned is quite similar to the purpose of this research.

3.3 Geneva Emotion Wheel

Emotions can be extremely hard to define even though they are used in everyday language. When it comes to defining the terms "emotion" or "feeling", it is the exact ambiguity of the definition which makes it hard to measure it. Emotions should be seen as dynamic processes in time rather than states [19]. When an individual reads an article, they will evaluate the consequences of the mentioned events and how they may affect his well-being and goals. The response triggered can be to adjust to the new situation or take action.

There are several ways in which a person can describe how or she feels. We have different approaches and emotion classification models to measure this. One of them is the Geneva Emotion Wheel, hereafter referred to as GEW [21] [22]. This Geneva Emotion Wheel can be found in Appendix A on page 29.
The Geneva Emotion Wheel makes it easier for a person to report the type of emotion he or she experienced while reading an article. A person will first identify approximately what the event meant to them and then pick the one adjective that best corresponds to the kind of feeling they felt. Twenty different emotions are arranged in a circle, the words may refer to a whole range of similar emotions i.e. anger also covers mad or being cross.

The different emotions can be seen as families that are arranged in a wheel shape with the axes representing two major appraisal dimensions i.e. control and valence. These two dimensions are further differentiated by arousal level. The smaller circles are meant to indicate the strength of the emotion. However, in this research the strength and two adjectives of every quadrant were left out, so the scope of the study would not be too broad. The adjusted Geneva Emotion wheel can be found in Appendix B on page 29.

There are several other existing measurement tools such as Self Assessment Manikin [2], which is completely non-verbal and uses images. Eleven methods of evaluation have been analysed by Janine Beker ([6]) and both the GEW and the Self Assessment Kin were found to satisfy industrial requirements. But in Natural Language Processing we found the GEW to be a lot more discrete and the terms used correspond to the way one naturally talks about emotions. Due to this reason it was decided to use the GEW as a tool to assess emotions.

4 Experiments

In this section we will describe how the experiments on the data have been conducted and what approach we took in order to be able to conclude whether it is possible to predict the emotion of a person after reading a news article.

4.1 Data collection

The data used in this research was collected by a group of colleagues, or alumni, of Leiden University. It has been collected from the websites of well-known Dutch news channels, the names of these will not be disclosed for confidentiality reasons. In this research only data of one news channel was used, as cleaning it and transforming it to the right format was time-consuming. Furthermore, labelling data from every news channel separately would have taken up more time than is intended for the project. For the labelling of data from every news channel we would also need more volunteers than we had.

4.2 Pre-processing

When it comes to Machine Learning, data pre-processing seems to be an overlooked topic, even though it could cause less accurate results. For consistent results the data used as input has to be in a certain format. The collected data has to be transformed or pre-processed by removing noise or rather unnecessary markup and metadata, text file headers and footers. The code used to remove the aforementioned noise is shown in Appendix D on page 31.

We will step-by-step explain what is being executed in the code:
1. Special characters are usually non-alphanumeric characters, numeric characters or symbols such as @. They do not have any added value to an emotion thus causing extra noise that is removed.
2. All the single character words consisting of just one character are removed.
3. There are cases in which incorrect data extraction leads to multiple spaces between words where only a single space is necessary. We in this step then replace all multiple spaces with a single space
4. To continue, all the words in the dataset or collection of texts (herein after referred to as corpus) were transformed into lower case, because "house" and "House" are essentially the same word in this text source and do not illustrate a difference in emotion a person would feel while reading. When it comes to social media this would have been different as a tweet in all caps makes it look like the person is shouting rather than simply putting their opinion or news out there.
5. Lemmatization is applied to make sure "cats" and "cat" are seen as the same word. Plural words are transformed to their singular form, which is necessary as a plural word does not increase or decrease the feelings a person already has. Reading cats or cat has the same impact.
6. The documents are converted to a matrix of tokens. While doing this all the words that are commonly used e.g. "het", "de" and "een" are called stopwords. Usually it is words such as "the" and "an" but because our articles are in Dutch we had to use the Dutch library of stopwords. The Python library NLTK has a list of such words. In news articles these words occur most often, in the model that has to be built these would thus take up valuable processing

time which is why they are removed.
After receiving the data, we also noticed there were quite a few empty files or rather files consisting of just headlines and not an actual article. Due to this we have to manually navigate through all the articles quickly to remove those and only send out a full fledged article to the reviewers.

4.2.1 Removing numbers

When analyzing the words that caused an impact, it was noted that there were still numbers present in the Bag of Words. It seems the code for removing special characters did not remove numbers thus we had to add a few more lines to improve the code and results.

Removing numbers in the pre-processing caused an increase of 8 percent in the accuracy of the Random Forest Algorithm results. The fact a number is irrelevant and should generally not be a determining factor of an emotion, removing it was essential and caused a significant improvement.

4.3 TFIDF

The data we have is in textual format and has been converted to a numerical format. This is a method or formula often used in information retrieval and text mining. By making use of this method we evaluate how important a word is to a document in a corpus. The importance increases proportionally to the amount of times a word appear in a document but it is offset by the frequency of the word in the entire corpus. This has been done using a TF-IDF matrix. This TF-IDF matrix makes use of the formula:

$$tfidf = tf * idf \tag{1}$$

$$idf(t) = \log \frac{n+1}{df(d,t)+1} + 1 \tag{2}$$

The terms used here are n, tf and idf, which can be described as following:
- N stands for the total amount of documents - TF stands for term frequency (how many times a word appears in a document). With CountVectorizer in the Python code above we count the number of words. Every article is different, some are long and some are short, due to which a term could possibly occur more often in a long document than a short one.
- IDF stands for inverse document frequency (giving more importance/weight to a rare word appearing in the documents, rather than the common ones). With calling "vectorizer.fit" in line 2, we calculate this. Words such as "is" and "of" may appear many times but they are actually not very important. The weight for these frequent terms needs to be down while for the more rare terms it needs to be scaled up. The Python code to apply this formula on the corpus looks as following:

```
vectorizer = CountVectorizer(max_features=1000, min_df=5, max_df=0.7,
stop_words=stopwords.words('dutch'))
X = vectorizer.fit_transform(documents).toarray()
tfidfconverter = TfidfTransformer()
X = tfidfconverter.fit_transform(X).toarray()
```

In section 4.5 we analyze the important words in this research and how much of an influence they had on being classified into a specific category.

4.4 Data labelling

The labelling of data in a research is an extremely important part and shapes the rest of the project. In the cases of emotion classification it is one of the biggest challenges as well. It is necessary to be attentive because a small mistake can result in inaccuracy and negatively affect the quality of the results and the performance of the model. In this thesis, this has been avoided as much as possible by making use of hand-labelling and having volunteers belonging to different study backgrounds and carefully instructing them before handing out the articles. The instructions we provided have been been described in section 3. Also note that by definition this is a subjective task, different people will associate a different emotion with the same text.

The data has been labelled by 5 volunteers, including the supervisor and researcher herself due to time constraints. Hand-labelling done by volunteers is considered to be an expensive task by most researchers. Also note, we were thinking of extracting an amount of articles and getting the same article labelled by more than a single volunteer, so we could also delve into the agreement on emotion between different volunteers. However, this would result in less articles labelled overall so we decided to not go through with it and stick to more articles, each labelled by one volunteer. The 5 chosen volunteers belong to different backgrounds, increasing the unbiasedness of the labels. One of the volunteers works as a Sales Manager for an IT firm in the United States, one is an Economics and Human Resource Management student while the last one is a Human Resource Management studies student only.

These volunteers were only shown the adjectives, so they would not have to take into account whether their emotion is regarded as negative, positive, high control or low control. Limiting them to just the adjectives has the advantage of the volunteers not getting too confused taking too many factors into account. Especially for the Human Resource student the explanation of having to take into account the different intensities would have made their feelings or decision which one to choose more complicated. The option of none was added, but the volunteers were instructed to avoid picking this one so we would have as much information as possible. Picking blank (none) only in the cases where there is literally nothing they feel. The reviewers were presented with both the Dutch and English adjectives that we will describe in section 3.6.

4.5 Modelling

There are several algorithms that can be used to build a classifier that predicts the emotion associated with an article. In this research we made use of Naive Bayes, Random Forest and

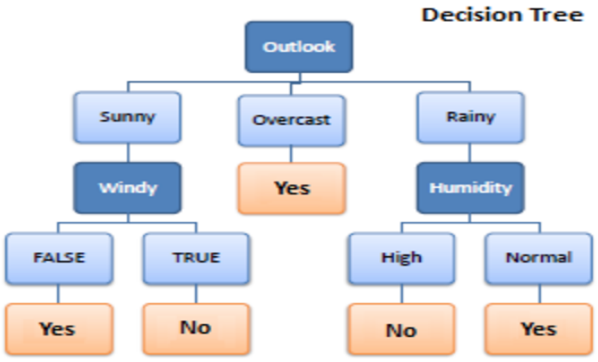


Figure 1: a decision tree on playing golf outside

Support Vector Machine (SVM).

4.5.1 Naive Bayes

Naive Bayes is an algorithm that uses the probabilities of each article belonging to a category to make a prediction. As it is making use of the BOW approach, whether the word sad appears at the start of an article or at the end is not important, what needs to be taken into account is how often it appears. It is a supervised learning approach that does not need a lot of data to perform well. Known to be an efficient and robust algorithm, especially for small sample sizes it has some drawbacks as well. Raschka [16] states "However, strong violations of the independence assumptions and non-linear classification problems can lead to very poor performances of Naive Bayes classifiers". Naive Bayes depends on the conditional independence assumption. The calculation of the different probabilities assumes that features are independent given the class, i.e.in our case keywords are conditionally independent given the article category. For example an article belonging to the "disappointment" category could be dependent on many words like "sad" and "cry" appearing in an article and the algorithm will assume these are not related to each other. It is called naive because the conditional independence assumption typically does not hold in practice. We do not have a reason to assume features are related, but at the same time it is wrong to assume the opposite. Even though this assumption is typically violated in practice, Naive Bayes can still have good performance, as its a robust and stable classifier with low variance.

4.5.2 Random Forest

Random forest is another supervised learning algorithm that is used for classification and regression. It is flexible and easy to use, which is why it is generally one of the most used algorithms. It builds a "forest", meaning it creates multiple decision trees from randomly selected training sets and combines the predictions of all trees into one single prediction. It is an algorithm that is designed to overcome the limitations of decision trees, as decision trees are prone to overfitting. In decision trees we have a tree-like structure with branches called nodes. It can be visualized as shown in figure 1.

A decision is made by traversing through these nodes, answering sequential questions. It is basically answering "if this, than that" conditions leading to a specific result. However, due to

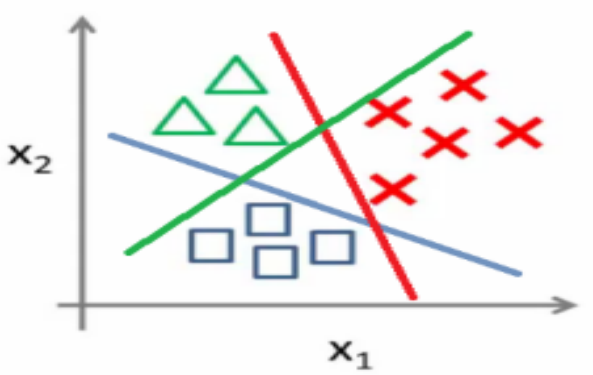


Figure 2: data projected onto a plane

the specificity within the tree it could lead to a very deep tree with many questions that may or may not even be relevant. Random Forest algorithm limits this by making use of multiple decision trees, in which we can utilize a few random features, if this is done for every tree we will eventually include most or at least many of our features. After every tree has made a prediction, these are all aggregated into one final result. Since in the end, all the trees errors cancel out when combined as different trees in the forest will overfit in different ways and thus voting averages these differences out.

4.5.3 Support Vector Machines

Support Vector Machine or SVM is another popular supervised machine learning algorithm for regression or classification. It is known to work well with text classification problems. SVM works with decision boundaries, it takes all the current labelled data and projects it onto a n-dimensional plane, with n the number of features you have. The value of each feature is the value of a particular coordinate. Usually it is used for binary classification but in our case we have many categories so a multi-classification problem. It can be visualized as shown in figure 2.

In the plane shown above we can interpret class 1 as sadness, class 2 as joy etc. having many different classes (12 to be specific in our case). Support vectors are the co-ordinates of individual observation. The classifier is a frontier that segregates the different classes. The hyperplane is a decision plane separating a set of objects having different class memberships. The main purpose is to segregate the dataset in the best possible way, so selecting the hyperplane with the maximum possible margin between the support vectors. Now it is obvious why it is used for linear problems, but for multi-class finding the maximum margin requires applying the "kernel trick" proposed by Aizerman in 1964 [15]. He explains how data that is not linearly separable in a n-dimensional space may be linearly separable in a higher dimensional space. It is not necessary to compute the exact transformation, the algorithm just finds the inner product of the data in the higher dimensional space. In our research we are using a linear kernel with a one-versus-one approach, which splits the 12 different classes (emotions) into one binary classification problem per each pair of classes.

English	Dutch
Anger	Boosheid
Contentment	Tevreden
Compassion	Medelijden
Disappointment	Teleurstelling
Disgust	Afkeer, Walging
Fear	Angst, Bang
Guilt	Spijt, Schuld
Interest	Interessant
Joy	Vreugde, Blijdschap
Pleasure	Genoegen, Plezier
Relief	Opluchting
Sadness	Verdrietig

Figure 3: Translation of the English words to Dutch

4.6 Language evaluation

As the news articles are in Dutch, these emotions have all been translated to Dutch. A study conducted by Nes, Deeg, Abma and Jonsson on the language differences in qualitative research explains how meaning can be lost during the translation process of one language to another [5].

Language differences can have consequences, because concepts in one language may be understood differently in another language. One of the reasons why it is difficult to translate from one language to another is polysemy. For example in English itself, the word plain can have different meanings, one means ordinary while another means easy. Firstly, you have to decide which sense of these words is needed. Once that is decided, you have to translate it to the most accurately representing word in Dutch.
In this study the meaning and interpretation of the twelve words; Anger, Compassion, Contentment, Disappointment, Disgust, Fear, Guilt, Interest, Joy, Pleasure, Relief and Sadness is central. So to minimize the effects of the risk of losing meaning there are different measures used. During the translation process of the words a peer-review was used and a discussion with speakers of both the languages; Dutch and English was organized to increase the validity. The definition of the words was evaluated to increase the accuracy of the translation. Anger translates to "boosheid" while sadness translated to "verdrietig". This was done taking our context in consideration. As anger in the sense we are using here literally refers to "boosheid". The translation that was decided on are shown in Figure 3.

As can be seen, for some terms we had to use more than one Dutch word to convey the actual meaning. We would also like to mention that since most of our reviewers had decent English skills, we decided to present them both the English and Dutch adjectives so the emotion would be even more obvious.

5 Results

In this section the outcome of the experiments, forming the basis of the thesis, will be presented.

5.1 Article distribution

Each of the volunteers was given a hundred articles. The articles have been extracted from the MongoDB database. MongoDB is a cross-platform document-oriented database program. Making use of the following SQL query we extracted the first 500 articles from the database:

```
1  mongo_uri = "mongodb://XX:XX@XXX.XXX.XXX.XX:XXX"
2  connection = MongoClient(mongo_uri)
3  db_name = "XXX"
4  db = connection[db_name]
5  db_coll = db["article-cleansed"]
6  mylist = list(db_coll.find({},{'cleantext':1}).limit(500))
7  count = 0;
8  for item in mylist:
9          count+=1
10         filename = '{}.txt'.format(count)
11         with open(filename,'w') as f_out:
12                 f_out.write('{}\n'.format(item))
```

The XX in the code are the account ID and password. The explanation of the code is as follows:
- In lines 1 to 5, we are connecting to the MongoDB database.
- In lines 7 and 9, we are creating a counter and increase it each time we encounter a new article. So we can count the amount.
- In lines 10 to 12, we are making sure to only get the first 500 from the database and save them in different text files in a new directory.

These 500 articles have been chosen to be labelled. They are randomly chosen, so there is a possibility that one volunteer might receive more sports related or political news articles. But as reading every article before handing it out is nearly impossible, this is something what we could not avoid. However, these articles all belong to the same time period of 4 or 5 months, they have been published in the same year so many of them contain the same news. This could mean that different volunteers could receive the same news published at a later time with an update. This is actually a positive factor as different volunteers could be labelling the same type of articles, decreasing the biased-ness as they might have different emotions after reading.

5.2 Emotion modelling

After receiving the labelling from the volunteers it was clear that there were about 5 to 10 articles of the 100 provided to every volunteer which evoked no emotion. The results of the labelling done by volunteers is graphically represented in figure 4.

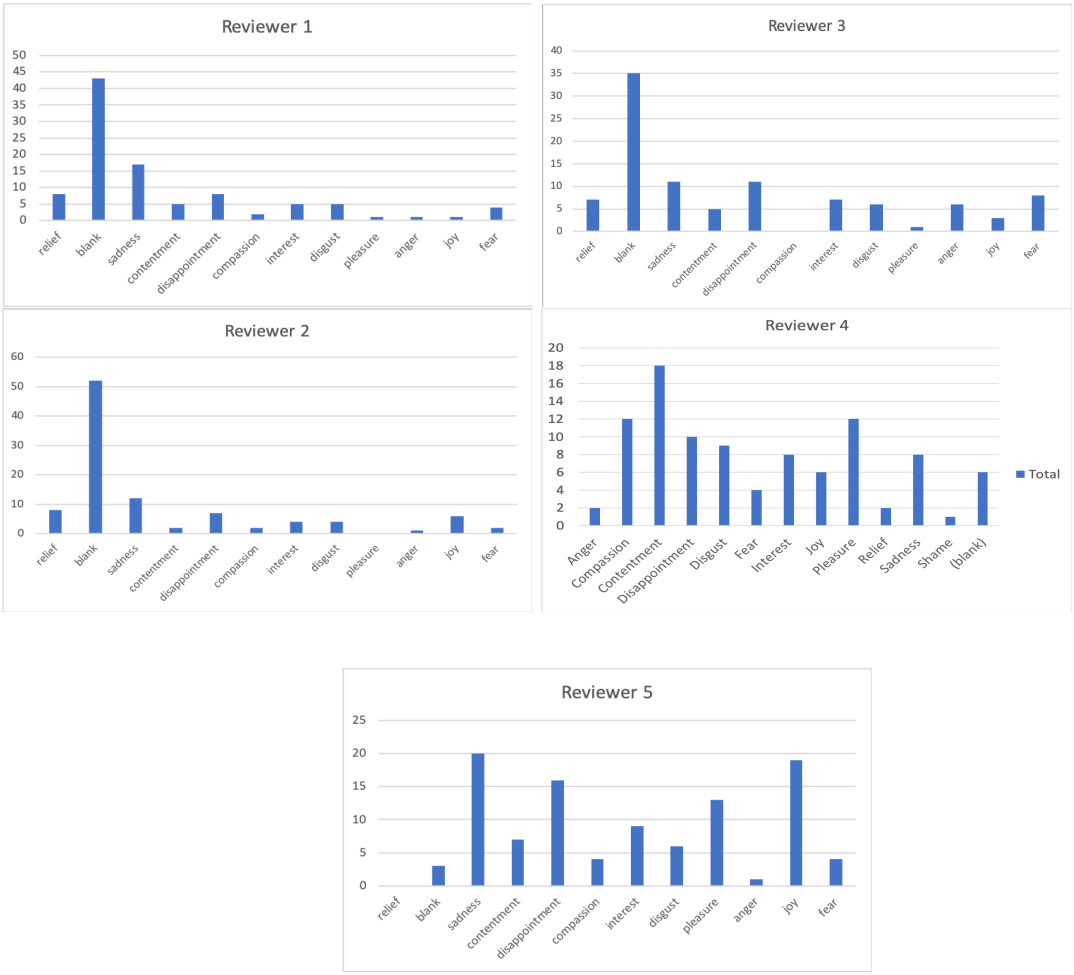


Figure 4: Results per volunteer

Most of the articles evoking no emotion consisted of announcements or sports results. Announcements or updates about sports results will generally not awaken any strong emotion, unless we have a sports fan as a volunteer. Also evident is that the Human Resource Management student, reviewer 5, had more articles that they could associate a feeling with, than the Computer Science students (reviewer 1 and 2). The graphs of reviewer 1, reviewer 2 and reviewer 3 show a similar trend with nearly none of them feeling "joy" or "pleasure" when compared to the remaining two reviewers. Some of these volunteers could be asked to re-label the articles that evoked no emotion further on. we elaborate on this in section 5.

For our research we decided to leave out the articles that evoked no emotion, as we want to focus on the media that does have an effect on readers. The complete table with the distribution can be found in Appendix F.

5.3 Tuning parameters

To predict the categories that the articles belonged to, different machine learning methods were used as explained earlier in 3.4. The results achieved with every algorithm and different test sizes and unique words have been illustrated in figure 5.

- The unique words there stands for the amount of different words taken into account. This is shown in Appendix D line 26. These have been set to 1000 or 1500. We only used these two

Unique Words	Test Size	Naive Bayesian	SVM	Random Forest
1000	20	29.9	29.9	41.3
1500	20	32.5	28.6	39.0
1000	30	31.0	31.0	35.3
1500	30	29.3	31.0	34.5

Figure 5: The accuracy score for each method with given parameters

Algorithm	Unique Words	Test Size	AUC
Naive Bayes	1500	20	0.63
SVM	1000	30	0.62
Random Forest	1000	20	0.67

Figure 6: The best AUC score for each algorithm

settings as more than 1500 seemed to make no significant impact on the results.
- The min_df was set to 5, meaning the words appearing in less than 5 documents will be ignored. This makes sense as it is less likely if a word is appearing this scarcely, it will have any effect on deciding the category (feeling) associated with that article.
- The max_df was set to 0.7, meaning the words appearing in more than 70% of the articles will also be ignored. This makes sense because if a word is appearing that often, it is not necessarily related to that specific article. We want to focus on the word that appear often enough, but not less than 5 or more than 70% of the total.

There has been a lot of debate on how to measure and analyse the efficiency of an algorithm. The two most used methods are accuracy, as used above, and AUC. AUC is an abbreviation for Area Under the Curve and is often used for two dimensional spaces. A macro-average computes the metric independently for each class and then takes the average, treating or giving equal weight to all classes (labels). A micro-average gives equal weight to each per-document classification decision. It treats the entire set of data as an aggregate result, and calculates 1 metric rather than k metrics that get averaged together. The micro average was preferable over macro as we have quite a few classes (e.g. Joy) that have more examples or labelled data than other classes. In section 4.2 we can see how skewed the distribution is, so we should not overfit to a single class. This is also reflected in the results that can be found in E.

The best result was found with the random forest algorithm, with a micro average, using 1000 unique words and a test size of 20%. In figure 6 we have included the best AUC scores for each algorithm with the respective test size and unique words.

Looking at the table we can see an amount of 1000 unique words works better than 1500 for 2 out of the 3 algorithms used. If we had implemented more than these three algorithms we could have concluded that this works for the majority, but more on that can be found in chapter 5 as there we discuss what could have improved our results.

Word	TFIDF	Word	TFIDF
team	0.453596	winnar	0.338273
jood	0.442290	belgisch	0.193652
pol	0.265693	levend	0.259589
duits	0.246481	gehaald	0.244952
verdacht	0.223330	pol	0.430392

Figure 7: Five interesting words with high TFIDF-scores

Anger	maart	opgelopen	opgenomen	opgepakt	opnieuw	oranje
Compassion	aangeboden	wit	wisten	wist	milieu	militair
Contentment	zware	schip	horst	haren	school	schrijft
Disappointment	aangeboden	februari	familie	leider	sprak	voorstel
Disgust	maart	groepen	robert	grens	ronde	rotterdam
Fear	aangeboden	onderweg	ongeveer	onze	ooit	oorlog
Guilt	aangeboden	probeerde	proberen	procent	proces	programma
Interest	gelegd	rest	vuurwerk	hart	harde	hard
Joy	aangeboden	niemand	nieuws	nk	noemt	noodhulp
Pleasure	aangeboden	nacht	naties	nauwelijks	negen	new
Relief	maart	ondanks	onderzoeken	onderzoekers	onduidelijk	ongeveer
Sadness	aangeboden	landbouw	break	kwijt	kwartfinales	kritiek

Figure 8: The six most informative words per emotion

5.4 Importance of words

The words that appeared in the articles were also analyzed. There were certain words which had more of an impact on deciding which category an article was designated into. This was done using feature importance. Every word appearing in the documents can be seen as a feature, featurelogprob is a function part of the scikit-learn library which calculates the empirical log probability of features given a class. Appendix C has a top 25 of words with scores illustrating the importance of those words in deciding the category it eventually was classified into. There are a few strange words such as xff (6) and xafti(16), this could be due to the stemming described in 3.2 or the xml tags still being present in the data. The more unique a word is to a document, the higher the TFIDF score. In Figure 7 we handpicked some words from the top 200 highest TFIDF scores, which were particularly interesting.
While some words in the results were not very meaningful, the appearance of words such as "jood" or "belgisch" was a surprising factor. These words signifying racial backgrounds were interesting in how they seem to influence the reader to choose a specific category. This again signifies the unconscious biased-ness that the reviewers might be exhibiting while reading the articles. In Figure 8 we present the six most important words for every category in a table.

From the words shown, there are some words which are very obvious; like "noodhulp" for joy. "Noodhulp" stands for emergency aid, when this is provided to the needed it is always a joy or positive feeling. The word "aangeboden" is one which appears as the most important word for many categories, this could be explained by the fact aangeboden can have a happy/positive connotation as well as a negative or sad one. The word kwartfinales for sadness means there

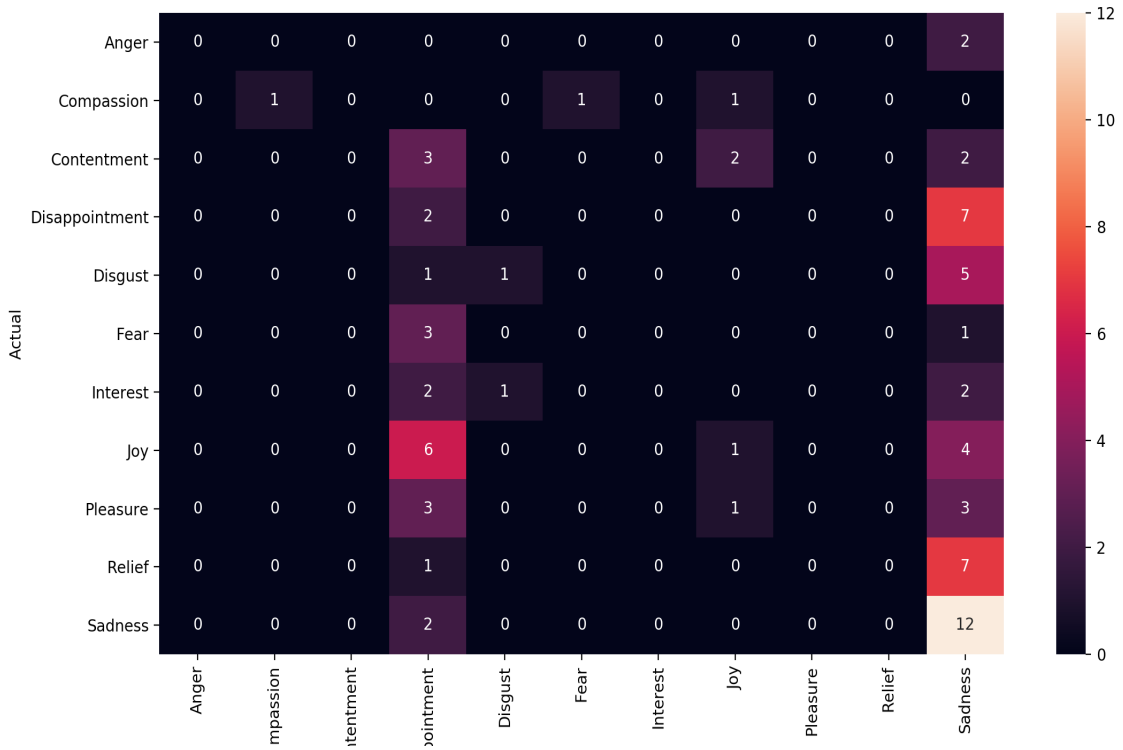


Figure 9: Confusion Matrix for the algorithm: Random Forest

were most probably articles where a "kwartfinale" was lost. Some words such as months of a year (e.g. februari and maart) have no explicit explanation. From the category compassion we can see that many people have a compassionate view towards "milieu", which stands for environment or climate, and the "military" protecting our country.

5.5 Correlation analysis

To understand why some articles were classified into certain categories we also analyzed the confusion matrices for the machine learning algorithms.
In figure 9, the x-axis represents the predicted class, while the y-axis represents the actual class. The class guilt was left out as there was no predicted label for that particular class and leaving it out made the confusion matrix a lot more readable. The ideal results are the ones where the predicted label is equal to the actual label. This is represented by the diagonal. There are are a few aspects that we noticed:

- The class disappointment, joy and sadness had the highest amount of labelled articles. As we can see in the confusion matrix, these are also the classes which have the highest amount of correct predictions. Going by this we can assume that having more articles for the other classes could have resulted in more accurate predictions for them as well.
- For some reason joy and disappointment were confused with each other, these might be sports related articles where one of the volunteers was filled with joy while the other was disappointed.
- Sadness and disappointment are also two classes intertwined. This is not surprising as these

two words are also closely linked. Disappointment results in sadness or the other way around is also often true.

The confusion matrices for the other two algorithms; Naive Bayes and SVM can be found in Appendix G and H respectively. The resulting confusion matrices for these algorithms were more or less the same. Naive Bayes did however perform better than the others for the majority class; sadness.

6 Discussion

We have been able to present the obstacles and difficulties with developing a machine learning model for such a complex problem. In this section, I would like to elaborate on how this research could possibly be used in the context of Bursting the Bubble and for other purposes.

6.1 The perfect machine learning model

If the research question was not this complex and we were able to come up with a nearly perfect model, having a high accuracy of 90 % or more, meaning we were able to predict the emotions associated with the news articles today. We could use this to decrease the polarization by analyzing which class and where this polarization is mostly coming from. If an article with Donald Trump was mostly classified into the disgust class, it means that Donald Trump is mostly being presented in a conventionally "negative" way. This could be depolarized by media releasing articles also from a different perspective, such as an article stating how much the unemployment rate has decreased thanks to him. This is a way of balaching the news, so the public can come up with their own opinions after reading many perspectives on the same topic or person.

6.2 The possible negative consequences

The results in this research could also be used for different purposes. If we look at a recent example, news concerning the outbreak of Corona virus was awakening a certain fear in people [20]. If our algorithm actually confirmed this fear among the public after reading news about COVID-19, many news agencies could misuse this fear and release more panic inducing articles in order to attract more readers. In fact, many agencies indeed started noticing this interest due to which more and more, supposedly panic inducing news was being published. The increasing abject panic and frustration to this news made the pandemic even worse. The public started buying so many paper products, causing a national shortage of toilet paper, followed by paper towels and facial tissues.

7 Future work

Several aspects has not been elaborated on due to the scope of the research. In this section we will describe these ideas that we did not delve deeper into.

7.1 Labelling

The labelling in this research was done by five volunteers. Even though the volunteers belonged to different academic backgrounds, all of them could have shown some biased-ness while labelling. To improve this the labelling can be done in a more unbiased way:

- If it is done by more volunteers, the articles will be read and exposed to more people evoking more and possibly different emotions. If there are different emotions evoked by the same article, the category chosen by the majority of volunteers can be chosen. Or we could study and compare these different emotions to each other as mentioned in 3.4.
- If there is not a possibility of more volunteers, the same volunteers could label a different set of articles from the 500 that we pre-processed, than the ones they were assigned to. That way every article could have been labelled at least twice improving the algorithm's capability of deciding which category it belongs to.
- The data that was collected for this research consisted of more than just one news channel. If the data for the other news channels could also be taken into account, we would have covered a bigger proportion of the current media. And eventually, come closer to the aim of the bigger research, Bursting the Bubble.

Using these approaches could result into more organized data, thus in better results.

7.2 K-nearest neighbour

Currently the machine learning methods used to determine are Random Forest, Naive Bayes and SVM. Though these are the most used methods for Text Classification tasks, there are many more like k-nearest neighbours which uses clusters to determine the label. It can be seen a method that applies the saying: "You are who you surround yourself with". It is a supervised learning algorithm and widely disposable in real-life scenarios and used for classification as well as regression. Taking a bunch of labelled articles, it learns how to label the remaining ones like every algorithm. The output is calculated as the class with the highest frequency from the k-most similar instances. Every instance votes for their class and the class which gets the most votes is taken as the prediction. Looking at figure 10 we will give a small example and how it is applied on tasks with two categories.

If you take the orange block marked with a ? as point P, of which you want to predict the label, the nearest neighbour or "k closest points" are found. If we take k as 3, we would take the three closest red stars and three green triangles. Then find the three closest point overall, which in this case is one red star and two green triangles. All three of these points vote for their own classes, since we have more green triangles in the three closest points the label attached to the triangles (B) will also be attached to point P. Thus P will also be classified into class B.

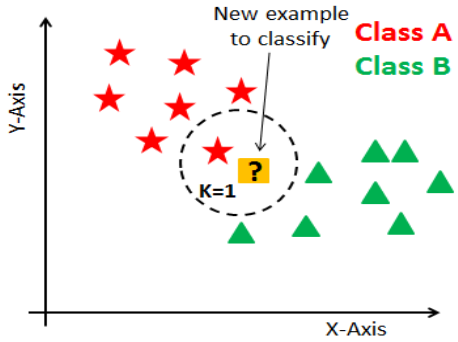


Figure 10: K-NN diagram with two classes

With this approach it is clear on which basis the specific cases or documents have been classified. The classification here is mainly driven on the basis of similar instances. As it is not only the predictability that matters, but also the transparency of the algorithm itself. By removing this "black-box" we will be able to show the same degree of trust in the model, as we previously had with traditional deterministic systems [23].

7.3 Semi-supervised learning

Semi supervised learning is an approach which also makes use of a big amount of unlabeled data and small amounts of labeled data. It is especially convenient when you do not have as much labeled data like in our case. As the name suggests it is halfway between supervised and unsupervised learning [4]. In this research if the labels can be predicted with a variable indicating the amount of certainty, a treshold could be applied to these labels. If the certainty, indicated in percentage, is higher than the treshold, the label can be attached to the article. This process is applied repeatedly on the data set using the continuously increasing new labelled data, to further label the remaining unlabelled data. Furthermore, the newly attained labelled data through this process can be used to train other models without the expensive task of hand-labelling.

8 Conclusion

The research question of this thesis reads: "To what extent can we predict the emotion of a person after reading a Dutch news article?". The answer to this will be deducted from the results presented in sections 3 and 4.

A few insights and findings have been revealed in this research. There was data available from a single news channel, not enough volunteers for the hand-labelling and time constraints in which it had to be completed. When looking into the models, the model providing the highest accuracy was Random Forest. Considering the complexity of the research, the expectations were not very high to begin with, despite that it was noticed that certain words seemed to have a significant impact on the readers and their emotions.

It was found that having more labelled data per class could would have resulted in a less skewed distribution and possibly more satisfying and reliable results. While it was interesting to see how some words seemed to influence the reader into choosing a particular category, considering the overall accuracy rate and AUC score, for now it is difficult to say how reliable it is.

All in all, the methodology and models give some insight into the emotions associated with the articles. However, with the current results and limitations, taking the best performing model into account with an accuracy score of 41.3% it can not be concluded whether machine learning can be used to predict the sentiment or emotion of a person when reading a news article.

References

[1] Gingerich, J. (2017). Print Media More Trustworthy than Digital. Retrieved 13 June 2020, from https://www.odwyerpr.com/story/public/9862/2017-12-08/print-media-more-trustworthy-than-digital.html

[2] Bynion, Teah-Marie Feldner, Matthew. (2017). Self-Assessment Manikin. 10.1007/978-3-319-28099-877−1.

[3] Taj, Soonh Fatemah Meghji, Areej Bakhtawer Shaikh, Baby. (2019). Sentiment Analysis of News Articles: A Lexicon based Approach.

[4] Chapelle, B. Schlkopf, and A. Zien, Eds. (2006) Semi-Supervised Learning. 978-0-262-03358-9.

[5] van Nes F, Abma T, Jonsson H, Deeg D. Language differences in qualitative research: is meaning lost in translation?. Eur J Ageing. 2010;7(4):313316. doi:10.1007/s10433-010-0168-y

[6] Benker, J. (2011). Incremental Analysis of Affective Evaluation Methods in the Context of Industrial Requirements. University College London.

[7] K. Shriniwas Doddi, Y. V. Haribhakta, P. Kulkarni (2014). "Sentiment Classification of News Articles", International Journal of Computer Science and Information Technologies, vol. 5, no. 3, pp. 4621-4623.

[8] Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. London: Viking/Penguin Press.

[9] Panke, Stefanie Stephens, John. (2018). Beyond the Echo Chamber: Pedagogical Tools for Civic Engagement Discourse and Reflection.. Educational Technology Society. 21. 248-263.

[10] Bruns, A. (2019). Are filter bubbles real?.

[11] Jack, Rachael Garrod, Oliver Yu, Hui Caldara, Roberto Schyns, Philippe. (2012). Facial expressions of emotion are not culturally universal. Proceedings of the National Academy of Sciences of the United States of America. 109. 7241-4. 10.1073/pnas.1200155109.

[12] Ekman, P. (1970). Universal Facial Expressions of Emotions. California Mental Health Research Digest, 8(4), 151-158.

[13] Russell, James (1980). "A circumplex model of affect". Journal of Personality and Social Psychology. 39 (6): 11611178.

[14] Arora, Deepali Li, K.F. Neville, Stephen. (2015). Consumers' Sentiment Analysis of Popular Phone Brands and Operating System Preference Using Twitter Data: A Feasibility Study. 2015. 680-686. 10.1109/AINA.2015.253.

[15] Aizerman, Mark A.; Braverman, Emmanuel M. Rozonoer, Lev I. (1964). "Theoretical foundations of the potential function method in pattern recognition learning". Automation and Remote Control. 25: 821837.

[16] Raschka, S. (2014). Naive bayes and text classification i-introduction and theory. arXiv preprint arXiv:1410.5329.

[17] Scherer, K. R. (2005). What are emotions? And how can they be measured? Social Science Information, 44(4), 693-727

[18] Davies, W. (2019). Why cant we agree on whats true any more?. Retrieved 13 June 2020, from https://www.theguardian.com/media/2019/sep/19/why-cant-we-agree-on-whats-true-anymore

[19] P. Verduyn and S. Lavrijsen, "Which emotions last longest and why: The role of event importance and rumination", Motivation Emotion, vol. 39, no. 1, pp. 119-127, 2015

[20] Mckeever, A. (2020). Coronavirus is spreading panic. Heres the science behind why. Retrieved 13 June 2020, from https://www.nationalgeographic.com/history/reference/modern-history/why-we-evolved-to-feel-panic-anxiety/

[21] Scherer, K. R. (2005).What are emotions? And how can they be measured?Social Science Information, 44(4), 693-727.

[22] Scherer, K.R., Shuman, V., Fontaine, J.R.J, Soriano, C. (2013). The GRID meets the Wheel: Assessing emotional feeling via self-report. In Johnny R.J. Fontaine, Klaus R. Scherer C. Soriano (Eds.), Components of Emotional Meaning: A sourcebook(pp. 281-298). Oxford: Oxford University Press.

[23] The Importance of Transparency in Machine Learning Models. (2020). Retrieved 14 June 2020, from https://medium.com/@perceptilabs/the-importance-of-transparency-in-machine-learning-models-368e16f360bc

Appendices

A

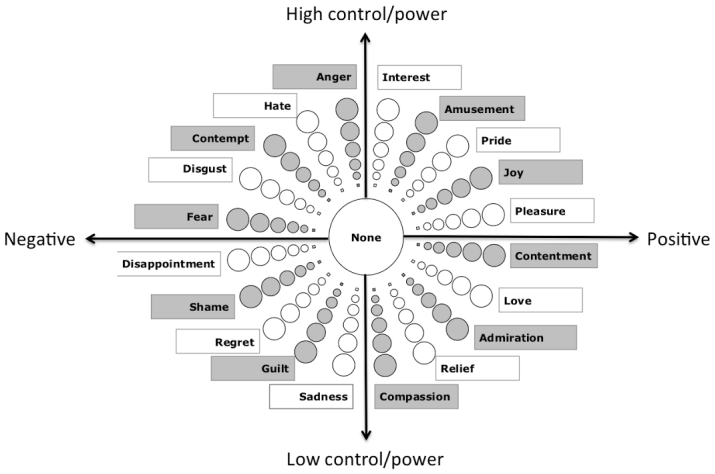


Figure 3: Geneva Emotion Wheel

B

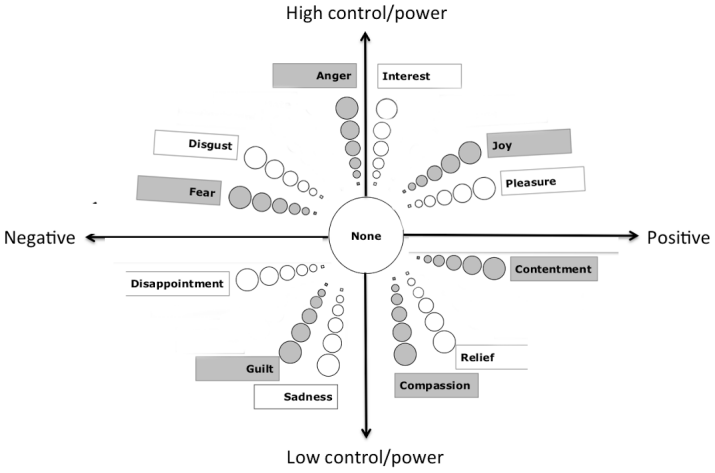


Figure 4: Adjusted Geneva Emotion Wheel

C

Rank	Unique Words	TF-IDF
1	wa	0.039389
2	jar	0.038745
3	dor	0.034975
4	mar	0.033040
5	nar	0.030732
6	xff	0.030352
7	hebb	0.029882
8	nederland	0.028953
9	mer	0.028636
10	word	0.027809
11	ha	0.026727
12	mens	0.026506
13	eerst	0.024927
14	war	0.023406
15	volgen	0.023245
16	xafti	0.022863
17	vel	0.022860
18	har	0.021949
19	twee	0.021693
20	teg	0.020528
21	geen	0.020477
22	twed	0.020284
23	amerikan	0.019057
24	kwam	0.018750
25	vorig	0.018659

Unique Words	Test Size	Naive Bayesian	SVM	Random Forest
1000	20	0.53	0.55	0.59
1500	20	0.54	0.54	0.62
1000	30	0.54	0.55	0.59
1500	30	0.54	0.55	0.59

Unique Words	Test Size	Naive Bayesian	SVM	Random Forest
1000	20	0.61	0.61	0.67
1500	20	0.63	0.61	0.66
1000	30	0.62	0.62	0.64
1500	30	0.61	0.62	0.64

D

```
for sen in range(0, len(X)):
    # Remove all the special characters
    document = re.sub(r'\W', ' ', str(X[sen]))

    #remove numbers
    document = re.sub("\d+", "", document)

    # remove all single characters
    document = re.sub(r'\s+[a-zA-Z]\s+', ' ', document)

    # Substituting multiple spaces with single space
    document = re.sub(r'\s+', ' ', document, flags=re.I)

    # Converting to Lowercase
    document = document.lower()

    # Lemmatization cats = cat
    document = document.split()
    document = [lemma.lemmatize(word) for word in document]
    #document = [stemmer.stem(word) for word in document]
    document = ' '.join(document)

    documents.append(document)

from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(max_features=1000, min_df=5, max_df=0.7, stop_
X = vectorizer.fit_transform(documents).toarray()
```

E

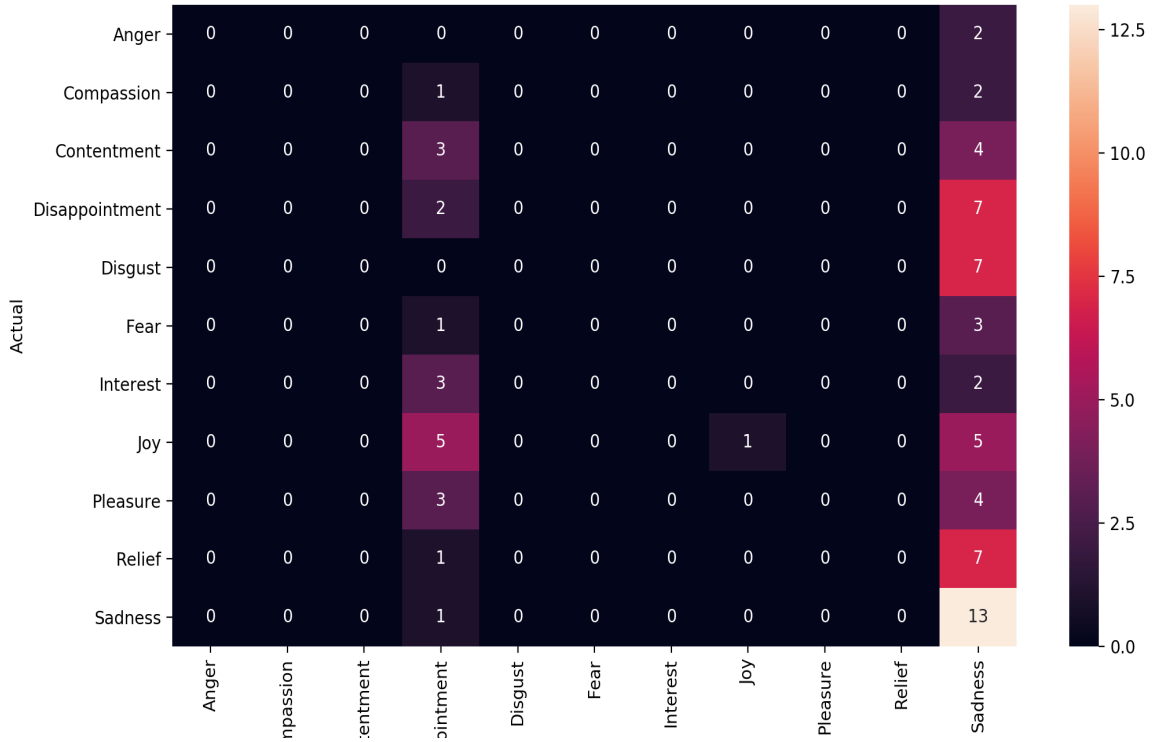
The complete table of AUC scores using a macro average
The complete table of AUC scores using a micro average

F

	*Reviewer*1	*Reviewer*4	*Reviewer*2	*Reviewer*3	*Reviewer*5	*Total*
Anger	1	2	2	6	1	12
Contentment	2	12	2	0	4	20
Compassion	5	18	2	5	6	36
Disappointment	8	10	7	16	16	57
Disgust	5	9	4	8	6	32
Fear	4	4	2	8	4	22
Guilt	0	1	0	2	0	3
Interest	5	8	5	7	9	34
Joy	3	6	9	4	19	41
Pleasure	1	12	2	1	13	29
Relief	8	2	8	7	2	27
Sadness	17	8	16	21	10	72

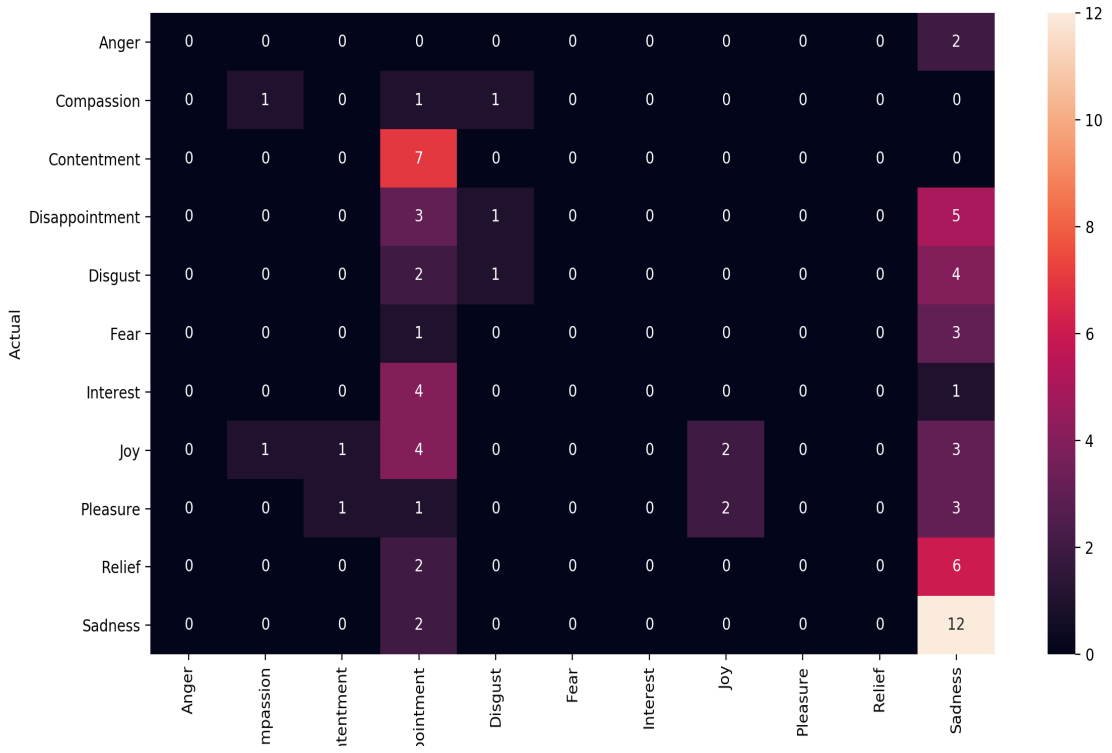
G

Confusion matrix for Naive Bayes



H

Confusion matrix for Support Vector Machines


Actual
Anger
Compassion
Contentment
Disappointment
Disgust
Fear
Interest
Joy
Pleasure
Relief
Sadness
0 0 0 0 0 0 0 0 0 0 2
0 1 0 1 1 0 0 0 0 0 0
0 0 0 7 0 0 0 0 0 0 0
0 0 0 3 1 0 0 0 0 0 5
0 0 0 2 1 0 0 0 0 0 4
0 0 0 1 0 0 0 0 0 0 3
0 0 0 4 0 0 0 0 0 0 1
0 1 1 4 0 0 0 2 0 0 3
0 0 1 1 0 0 0 2 0 0 3
0 0 0 2 0 0 0 0 0 0 6
0 0 0 2 0 0 0 0 0 0 12
Anger
Disgust
Fear
Interest
Joy
Pleasure
Relief
Sadness
12
10
8
6
4
2
0