Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

Emiil Kløvvik

Determining the age and gender of an individual based on text classification

Comparing two binary classifications with one 4-class classification

Master's thesis in Information Security
Supervisor: Patrick Bours
Co-supervisor: Muhammad Ali Fauzi
December 2021

NTNU
Norwegian University of
Science and Technology

Emiil Kløvvik

Determining the age and gender of an individual based on text classification

Comparing two binary classifications with one 4-class classification

Master's thesis in Information Security
Supervisor: Patrick Bours
Co-supervisor: Muhammad Ali Fauzi
December 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

Abstract

Age and gender detection is one of the tools that can be used to provide a form of safety in chatrooms. By finding the correct age group of an author of a chat, or text, this study could protect young children, either from posing as young adults online themselves or from predators seeking them out, pretending to be children. This study seeks to improve the detection of age and gender through text classification by finding the differences between looking at age and gender classification as two separate binary problems, or as one 4-class classification problem.

By running six different algorithms, three different feature extraction methods, and implementing soft and hard voting on the results, from both the binary classifications and 4-class classifications, it provides a solid basis for comparison. The metrics chosen as comparative numbers are accuracy, precision, recall, computing time, as well as $F_{0.5}$ and F_1 scores. The focus is on precision and the $F_{0.5}$ score because, given the potential application in detecting predators, it is more relevant to detect adults posing as children. This is given that the classifications for the binary methods are based on a child being class 1, and an adult being class 0. The results from the 4-class classification are also combined into two parts, one for age and one for gender, in order to have more comparable results.

Intermediate results show that hard voting has a more substantial effect on the results than soft voting. It does so for both the binary and the 4-class combined data, but mostly for the 4-class classifications.

The results show that the computing time for the 4-class classification is by far the faster choice, as the classification for the binary data must be run twice. The differences with regards to the other metrics vary between the different methods and range from negligible to 60%, where the highest differences occur for the worst performing methods overall, on gender classification and hard voting. The difference in average precision and $F_{0.5}$ score is 1.6% and 4% respectively, in favor of the 4-class combined data classification. Looking at specific authors, and if the classification differed between binary and 4-class combined classification, the latter classifies 4.3% more authors correctly.

The difference between the different methods is not always significant, but from an overall standpoint, the 4-class combined data classifications perform better in 70.8% of the methods used in this study, with regards to precision and $F_{0.5}$ scores. This suggests that this approach could be the better choice in detecting age and gender through text classification in e.g., chatrooms.

Sammendrag

Alder og kjønndeteksjon er en av verktøyene som kan brukes for å sørge for en form for sikkerhet i chatterom. Ved å finne riktig aldersgruppe på en bruker ved hjelp av teksten den har skrevet, kan denne studien beskytte unge barn, både fra å utgi seg som unge voksne på nettet, og fra overgripere som utgir seg for å være barn. Denne studien vil forsøke å forbedre deteksjon av alder og kjønn ved tekstklassifisering, dette vil gjøres ved å finne forkjeller mellom å se på alder og kjønnklassifisering som to separate binære problemer, og et 4-klasse klassifiseringsproblem.

Ved å bruke seks forskjellige algoritmer, tre forskjellige måter å hente attributter på, og implementering av to forskjellige måter å behandle resultatene, for både binær og 4-klasse-klassifisering, sørger studien for et solid grunnlag for sammenligning. Beregningene som er valgt til å brukes i sammenligningen er *accuracy*, *precision*, *recall*, databehandlingstid, i tillegg til $F_{0.5}$ og F_1 score. Fokuset vil ligge på *precision* og $F_{0.5}$ score, ettersom det er et potensiale for å bruke dette til å detektere overgripere, vil det være mer relevant å detektere voksne som utgir seg for å være barn. Dette er basert på at klassifiseringen for de binære metodene klassifiserer barn som 1 og voksne som 0. Resultatene fra 4-klasse-klassifisering blir også kombinert til to deler, en for alder og en for kjønn, slik at resultatene blir sammenlignbare.

Mellomliggende resultater viser at *hard voting* har en større påvirkning på resultatene enn *soft voting*. Dette gjelder både for binær- og kombinert 4-klasse-klassifiseringer, men mest for 4-klasse-klassifiseringer.

Resultatene viser at databehandlingstiden til 4-klasse-klassifisering er markant raskere enn for to binære klassifiseringer, ettersom de må kjøres to ganger. Forskjellene vedrørende de andre beregningene varierer mellom de forskjellige metodene, fra omtrent ingen forskjell til 60%, hvor de største forskjellene skjer ved de metodene som samlet har dårligst resultater, på kjønnklassifisering med *hard voting*. Forskjellene i gjennomsnittlig *precision* og $F_{0.5}$ score er 1.6% og 4% henholdsvis, til fordel for kombinert data 4-klasse-klassifisering. Ved å se på spesifikke brukere, og om

klassifiseringen med binære og kombinert data 4-klasse-klassifisering er forskjellig, så klassifiserer sistnevnte 4.3% flere brukere korrekt.

Forskjellene mellom de forskjellige methodene er ikke alltid signifikant, men fra et overordnet standpunkt klassifiserer kombinert data 4-klasse-klassifisering med bedre resultater i 70.8% av metodene brukt i denne studien, med tanke på *precision* og $F_{0.5}$ scores. Dette tyder på at denne tilnærmelsen kan være et bedre valg med tanke på alder og kjønnsdeteksjon ved tekstklassifisering i for eksempel chatterom.

Preface

This thesis is the final delivery of a master's degree in Information Security at the Norwegian University of Science and Technology (NTNU) in the faculty of Information Technology and Electrical Engineering. It was written from August to mid-December 2021, as a part of research regarding Chatroom security, which aims to address the problem of online predators and how they use technology to their advantage and anonymize themselves. The topic chosen for this thesis has been to further improve the methods, and how to analyze the results when using machine learning to classify an individual based on text into age and gender.

Emiil Kløvvik

Lillehammer, Monday 13th December 2021

Acknowledgements

I would like to thank Patrick Bours for being my supervisor, for all insights, information, and tips regarding both the topic at hand and the writing itself. I would also like to thank Muhammad Ali Fauzi for his role as co-supervisor, and especially his help with providing the initial classification results that formed the basis for my own analysis.

Also, a special thanks to two of my friends, Kais and Gard, who advised with their technical expertise, and my girlfriend Martine, who made sure I gave it my best.

Emiil Kløvvik

Lillehammer, Monday 13th December 2021

Table of Contents

Figures

Tables

Equations

Acronyms

ABD Author-Based Detection

BoW Bag of Words

CBD Conversation-Based Detection

CPU Central Processing Unit

FA Female Adult

FC Female Child

HV Hard Voting

k-NN k-Nearest Neighbors

LIWC Linguistic Inquiry and Word Count

LR Logistic Regression

MA Male Adult

MBD Message-Based Detection

MC Male Child

MCC Matthews Correlation Coefficient

NB Naïve Bayes

NLP Natural Language Processing

NN Neural Networks

NTNU Norwegian University of Science and Technology

PoS Part of Speech

PWC PricewaterhouseCoopers

RBF Radial Basis Function

RF Random Forest

RMDL Random Multimodel Deep Learning

SV Soft Voting

SVM Support Vector Machine

TF Term Frequency

TF-IDF Term Frequency Inverse Document Frequency

VSM Vector Space Model

$\boldsymbol{\chi^2}$ Chi-square

1. Introduction

1.1. Topic covered by the project

This study will look at a dataset composed of blogs written by different ages and genders and aims to explore a different approach to the classification of age and gender based on text. Past research has used the same methods in the form of an ensemble method with a specific set of machine learning algorithms and feature extraction methods, but only for either age or gender in the same classification. Here it will be done by running the same methods, but for a single 4-class classification including both age and gender. The results will not be calculated into one score using all the methods, as ensemble methods represent, but rather keep all the different scores from all the methods. By identifying potential differences between the two approaches, this research could help either improve the efficiency processing-wise, improve the detection rate, or the accuracy, or realize that this could be a less relevant approach to pursue.

While the study uses many machine learning algorithms; Logistic Regression, Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Neural Network, Decision Tree and Random Forest, its purpose is not to explain in detail how they work, but a general introduction will be given. The study will focus more on how the results from the classification could be used for further analysis. Some relevant related work will be introduced, whether it being research within predator detection or age and gender detection based on text or other forms of data. This should help provide the context enveloping this thesis, which is detecting predators in chat rooms and other online communication. The scope is to look at the sub-problem within that group, namely determining the age and gender of a “chatter”, or in this case a “blogger”. Finding out whether the chatter is a pedophile, or a predator is not within the scope, nor is the work of collecting data from real-life chats.

1.2. Keywords

Machine learning, Age and gender detection, 4-Class Classification, Multi-class classification, 2-Class Classification, Binary Classification, Chatroom security

1.3. Problem description

In the current field of author profiling, deception detection, and age and gender classification research, the focus seems to be on finding the best single method, or algorithm. The algorithm is chosen based on how well it performs with different feature extraction methods. Normally, when performing age and gender classification, either the gender or the age is determined first, then the other. This thesis aims to address both approaches in a way that the results from a selection of several methods being run as two binary classifiers: age (child/adult) and gender(male/female), and the results from one 4-class classifier (male adult/male child/female adult/female child), will be compared in order to find potential differences.

1.4. Justification, personal motivation, and benefits

This thesis is a part of the “chatroom security”-research at the Norwegian University of Science and Technology (NTNU). This research seeks to address the problem of pedophilia in chatrooms, specifically how the predators can portray themselves as children and manipulate children into doing their bidding. While this thesis indirectly aids the overall research, it does not focus on the problem itself but rather focuses on a part of the research´s goal, which is to differentiate between the different chatters, who they are, removing the cover of anonymity. One of the sub-problems of the overall research is determining their age and gender, regardless of what their profile or texts state. By researching this sub-problem, this thesis can help protect children, both from themselves if posing as young adults online or from predators that seek them out.

Predators vary in how they interact with their victims, from aggressively trying to exploit them from the beginning to building relations with them first. Episode: “Barnerov” (Amble, 2021) from season 2 of the series

“Norge bak fasaden”, with a guest appearance of this thesis’s supervisor Patrick Bours, mention the “Chatroom security” research, and reveals precisely how upfront the predators can be. While this is indeed one of the issues within this field, the victims themselves could be more affected by the experience if they are exploited by a person they trust, leaving them to struggle with their mental health for a more extended period. Other research focuses on finding the predators as fast as possible, e.g., within a few chat records, but this is not the focus here. The objective of this study is to improve the performance of classification of age and gender regardless of how long the chat is.

1.5. Research questions

The main research question that the thesis will try to answer is:

Does the classification of both age and gender through text analysis and machine learning differ when treating it as a 4-class problem rather than two binary problems?

The following sub-questions have been formulated to help answer the main question:

- Are there differences in computing time, by running the algorithms once with four classes rather than two times with two classes?
- In what way should the results from the classifications be processed and analyzed in order to achieve comparable data?
- What is the difference in performance on a per algorithm basis, using the 4-class classification and the binary classification?

The sub-questions are steppingstones needed for answering the research question and will provide a scope in order to analyze the work conducted in this study.

1.6. Planned contributions

The technical part of this thesis is of the more general kind, as most of the technical part is done outside its scope. Nevertheless, the primary stakeholders are the ones researching the same field and specifically those who develop detection of age and gender based on text, as this is meant to help improve that specific task. Other interested parties in this study could also be those working with the detection of predators other than being researchers for improving the field. Law enforcement as users of the detection, system administrators with regards to limiting access to specific content based on age or e.g., in my specific background in many years of digital forensics and incident response: author profiling within the detection of stolen email accounts. As this field progresses, so does the amount of data. Therefore, more research with regards to more efficient ways, either classification or performance-wise, to process this kind of data should be pursued, which is what this thesis seeks to do.

1.7. Reader guide

Chapter 2 describes the background for this thesis, including the state of the art of age and gender classification and related work that implements the different most popular machine learning algorithms for similar purposes. The chapter ends with an introduction to the chosen algorithms for this study. Chapter 3 presents the methodology chosen to be able to answer the research question, how the data was collected and how it will be analyzed. Chapter 4 details the dataset, both the initial dataset consisting of sentences, and the dataset with classification scores. Chapter 5 presents the results from the different methods and the comparison of the two different approaches, highlighting specific findings. Chapter 6 concludes this thesis and suggests possible future work.

2. Background and literature

2.1. Background

2.1.1. Focus

The area of predator detection is closely intertwined with other topics such as author profiling, age and gender detection, and grooming detection. There is a lot of work in all these areas, both nationally in Norway and internationally. Not all these areas require transcripts or datasets that contains activity from true predators and victims to be able to contribute to the field of predator detection. This section provides an overview of the state of the art in several of these fields, as they all contribute to understanding the current state of the "art."

Since the research of detecting predators in online environments began using machine learning, there have been a lot of different approaches. One way to look at the different approaches is to divide them into three subcategories by what kind of features they look at, lexical and behavioral and a combination of these. Another popular approach uses several methods in combination, which is also referred to as ensemble methods.

Firstly, there are the lexical features. These features can be extracted from the text itself, not including the meaning of the words. Bag of Words (BoW), which will be explained later in section 2.3, is the most common way to extract these kinds of features. Examples of such features could be, as (Bello, et al., 2020) uses, word length, number of syllables, and how many emojis are used, in a combination of using e.g., unigrams, bigrams, or trigrams.

Secondly, there are the behavioral features which are focus on how the author writes and is an attempt to represent the author's habitual traits. This could be how long sentences the authors normally write, how often, how fast, how correct, or e.g., how many questions they normally ask other people in chats.

As a third category, we have a combination of both lexical and behavioral features. This is a commonly favored approach, as it includes both and because behavioral characteristics rarely are found without looking at the linguistic elements. Therefore, there is no section in this thesis solely for papers exclusively looking at behavioral features.

The fourth category within the field is the ensemble method, which in newer research has proved quite useful. This basically uses several methods (classification algorithms and features) individually, then combines the results, for example using hard and soft voting, to get a result based on all the algorithms.

Other research within machine learning, not focused on using text for age and gender classification will also be mentioned in section 2.2, as they implicitly contribute to and give a more correct picture of the state of the art.

Table 1, at the end of section 2.2 gives an overview of the different main related works, including datasets used, some of the results they achieved, what methods they implemented and what year the research was published.

2.2. Related work

This section will provide an overview of the state of art, presenting related work that either applies to having used lexical or behavioral features or both, and ones that have used an ensemble method within predator detection or age and gender detection. Other research that has used neither but is still related by what kind of algorithms have been used will also be introduced to give a broader view of the field. Table 1 at the end of this section includes the main research mentioned, what dataset they used, important results they achieved, what year the research was done, and keywords related to what method was used.

2.2.1. Lexical and behavioral features

Closely related research carried out by (Kulsrud, 2019) utilized Natural Language Processing (NLP) and attempted to detect cyber grooming as early as possible during an online conversation. While several approaches were developed, conversation-based detection (CBD) achieved the best results. The other two approaches attempted were message-based detection (MBD) and author-based detection (ABD). The MBD tried to classify each message as predatory or non-predatory but was quickly dismissed, as the author obtained poor results due to similarities between the messages written by predators and non-predators. Given examples were messages such as "Good", "Hi" and "Cool". These could relate to this study as it could affect the performance of the classifiers used, but as this study only performs processing and analysis of the results from the algorithms and not on the dataset itself it can be considered moot. The CBD was based on the work of (Villatoro-Tello, et al., 2012), who implemented a two-stage classifier for detecting predators, achieving an F_1 score of 0.8734 on the same dataset. Firstly, the classifier tried to detect conversations that involved a predator, and secondly, it attempted to determine who was the predator and who was the victim. The last approach was author-based detection, which consisted of gathering all the messages from a single author and using all of them as a basis for determining if he, or she, was a predator or not. This was in a single, binary classification stage. (Kulsrud, 2019), as with many others, used the PAN dataset from 2012. According to (Inches & Crestani, 2012) they created the dataset with several hundred thousand conversations. They only included a few conversations that included a potential predator, but a lot of what could be referred to as false positives, which could be sexual conversations, or other conversations themed like those had by predators. While the realistic number of predators is very low, they increased it by including data from Perverted Justice's (PJ) website, which contains conversations where one party is a convicted sexual predator, and the other party a volunteer posing as an underage potential victim. This could potentially affect results as it is not 100% authentic. PAN is, according to (Bevendorff, et al., 2020):

"a series of scientific events and shared tasks on digital text forensics and stylometry" (Bevendorff, et al., 2020)

and is one of the main arenas when it comes to providing data meant for a range of different tasks within e.g., classification using machine learning. PAN also hosts different competitions for mentioned tasks. The datasets they provide have default tasks for each dataset, for example, author profiling or identification, plagiarism detection, and deception detection.

(Kulsrud, 2019) achieved varied results from the different methods. The computations relating to both CBD and ABD used k-fold cross-validation which, although more expensive computational-wise, was done in order to achieve better out-of-sample performance. The research also employed pre-filtering and pre-processing in order to mold the dataset into something more applicable for the chosen classification methods. While the MBD applied to single messages did not yield good results, it still worked as intended. Pre-filtering was done by removing conversations with only one author, short conversations due to difficulties achieving accurate classifications with minimal amounts of data, group conversations i.e., conversations with more than two chatters involved, and messages either filled with multiple concatenated special characters or no characters at all, as they did not provide any value to the algorithms. By filtering according to these criteria, over 80% of the original dataset was removed. The PAN dataset is in its original form not balanced with regards to the number of adults, children, non-predators, and predators. Still, the organization behind PAN has already pre-processed the data in some regards, leaving (Kulsrud, 2019) only to further pre-process it by replacing a set of special characters with whitespaces, removing all other data than alphanumeric characters and whitespaces, reducing all concatenated whitespaces into single whitespaces, converting all capital letters to lower case and removing stop words found in the Natural Language Toolkit. All testing was done with and without the pre-processing and it had a varied effect on the different methods, as CBD performed better without it and ABD performed better with pre-processing. The best results on the conversation segments were achieved with the CBD where (Kulsrud, 2019) managed to get an $F_{0.5}$ score of 0.893, in which 209 out of 254 of the predators were classified correctly and 20 non-predators were classified incorrectly. The dataset included 218702 unique authors after pre-filtering and pre-processing. (Kulsrud, 2019) strived to detect predators as early as possible, which could filter out a lot of predators and possibly the ones who affect the victims the most.

The numbers presented by (Kulsrud, 2019) are by no means representative of the whole field of research but show some of the possibilities within predator detection. (Silva, et al., 2020) based their work on the PAN2018 dataset and tried to classify age and gender based on semantic, lexical, and syntactic characteristics. Even though they proclaim it is:

"Possible to characterize both the age and gender of an author with an accuracy greater than 50%." (Silva, et al., 2020)

The results are still noteworthy, as they came in eighth place in the PAN2018 competition, the language was both English and Spanish, and more importantly, it was not a binary classification. They achieved these results by classifying the author's age into their 10-year age range, their twenties, and thirties. As with all research based on datasets, they are bound by their limitations if not mitigated. (Silva, et al., 2020) did not have the luxury to be able to balance the dataset, as there were too few authors in the 10-year age range to be found. More specifically they divided the age ranges into 13-17, 23-27, and 33-37. The classification itself was based on one of the methods used in this study, namely Random Forest (RF), and a performance estimate by means of 10-fold cross-validation, akin to (Kulsrud, 2019)'s research.

Often the features can be used interchangeably between lexical and behavioral, depending on what context they are used in. (Holbæk, 2019) focused on determining if the author of a text is underage, younger than the age of 18, or older than the age of 25, an adult. The results confirmed that one of the best approaches found for author profiling with regards to age is e.g., Support Vector Machine (SVM) in combination with Term Frequency Inverse Document Frequency (TF-IDF), Linguistic Inquiry and Word Count (LIWC), n-grams, and Part of Speech (PoS), and it also showed that it was indeed possible with the dataset used in this thesis, the Schler dataset, which will be introduced later in chapter 4. (Holbæk, 2019) also used three of the PAN datasets (PAN13, PAN14, PAN15) to confirm the findings on the most used datasets used by the field, including social media and Twitter data. The features used in that study were both lexical and behavioral, or stylistic and contextual as he describes it. The experiments were done on different corpora and the best result, as mentioned earlier, on the joint corpus experiment was with SVM Radial Basis Function (RBF) and achieved an F_1 score of 0.89.

(Hancock, et al., 2004) and (Newman, et al., 2003) present findings suggesting that people often, as an everyday occurrence, show deceitful behavior through for example, paralinguistic or linguistic cues. As this research area also encompasses product reviews and other online communities not necessarily related to predators, it accentuates the broad specter of the field. (Bond Jr & DePaulo, 2006) and (Ott, et al., 2011) emphasize the need for machine learning in this kind of detection as well, as they assert that humans are only 50% effective in detecting deception and achieve an accuracy detection rate of 90% of deceptive opinion spam. (Banerjee, et al., 2014) used keystroke patterns to detect deception or deceptive behavior. Their research shows that there are clear signs that can differentiate between deceptive and truthful writing. This substantiates the research of (McCornack, 1997) and (Vrij, et al., 2006), who showed that deceptive behavior, or lying, poses a cognitive burden, which (Vizer, et al., 2009) and (Epp, et al., 2011) later proved affected keystroke features. (Banerjee, et al., 2014) focused on features like pauses, revision rate, and writing speed or rate, using SVM and BoW for classifying truthful or deceitful writing, achieving an accuracy of 0.943 on data pertaining support of gay marriage. Their data was obtained through Amazon Mechanical Turk, where users were asked to write both truthful and deceitful messages on one of the three topics: Gay marriage, gun control, and restaurant reviews. As a byproduct, they found differences in the use of adverbs, verbs, function words, nouns, and adjectives. Specifically, as (Newman, et al., 2003) and (Ott, et al., 2011) mentions explicitly, there is a difference in the use of 1st person pronouns. The less frequent it is, the more it could be attributed to psychological distancing.

(Huisman, 2016) tried to further explore the research done by (Banerjee, et al., 2014), looking exclusively at the keystroke dynamics of a user, not the words themselves. While (Banerjee, et al., 2014) got results that would indicate this to be a worthy pursuit, and (McCornack, 1997), (Vrij, et al., 2006), and (Epp, et al., 2011) all point to what can only be interpreted as good results, (Huisman, 2016) achieved a 0.13 to 0.15 accuracy with k-nearest neighbors (k-NN) on both the dataset provided to him by PricewaterhouseCoopers (PWC), consisting of metadata from 30 users answering a survey, and (Banerjee, et al., 2014)'s dataset. This could suggest that the chosen features such as dwell time, flight time, typing speed rate, deletion rate, and pause rate, was either not enough, used incorrectly, or not applicable as features for this kind of detection.

As with some of the other articles referenced in this chapter, (Pendar, 2007) approached the challenge of detecting pedophiles in chats with the SVM and k-NN models. He achieved an F_1 score of 0.943 on the test data from PJ's website, consisting of actual predators and actors, or volunteers, portraying themselves as young underage victims. One of the more difficult challenges in this field of research is procuring or developing a realistic or authentic dataset. If the study trains the methods to detect gender or age, not considering the predator aspect, or if the datasets only include actors instead of actual predators, the results would most likely be less helpful for detecting predators in real chat environments. This does not entail that the research does not improve the detection rate, but realistically it would probably be better with actual real data. Although the chats were somewhat orchestrated, they introduced another challenge as it was indeed chats. The features were extracted using uni-, bi-, and trigrams and preprocessed using a custom stop-word list. While the Schler dataset used in this thesis consists of blogs, it still retains a more formal touch. Chats include a lot of words not typically seen in other textual data, such as terms lengthened to emphasize how the author feels, e.g., "yes", "yeeees", and "yeeeeeeees", rendering default stop-lists ineffectual.

(Borj & Bours, 2019) also based their work on the PAN12 dataset as part of the chatroom security research. Using NLP, linear SVM on 1-gram features, they achieved the best results with an F_1 score of 0.86. Regarding stop-words, they assumed that they would gain a better accuracy and F_1 score if they kept them in the dataset. This seems a natural assumption as (Pendar, 2007) emphasized the use of chat-specific words that altered the meaning of words, e.g., the length or long pauses using several blank spaces. Nevertheless, they achieved the best results excluding the stop words. The shift in this field of research regarding types of datasets, moving from more standard texts to more informal ones, is mentioned as a new challenge. This shift makes behavioral features more prominent, as more of the author's habits and personal characteristics shine through. As with all research, the quality of the dataset sets the baseline of how accurate and applicable it is.

Albeit not a momentous challenge for this thesis using age and gender only, studies focusing on predators and victims have a much harder time finding datasets clearly and accurately marked with specific classes consisting of actual predator and victim data.

(Peersman, et al., 2011) wrote one of the articles that researches a topic like the problem that this thesis will attempt to unfold. The article examines short texts, or chats in this case, originating from the online social networking site Netlog, where they try to predict the age and gender of the author using SVM and unigrams. They used several approaches to find the most informative features in chat data and used the Chi-square (χ^2) feature selection metric. This metric was used to select four different feature sets, consisting of 1000, 5000, 10000 and the 50000 features with the highest χ^2 value. The challenge mentioned earlier, regarding consecutive identical letters, was avoided, as they removed all the consecutive letters after three letters, as “yeees” and “yeeeeees” would be the same. This is one way to handle this specific challenge, at the cost of losing some of the behavioral characteristics of the author. Another topic they explored is how much data is required as a minimum to get usable results. The experiment was conducted three times, one with 10000 posts, one with 5000, and lastly with 1000 posts. Firstly, they discovered that the accuracy and F_1 score improved when dividing the two age classes by several years so that the age group 11-15 (min16) and the age group above 25 (plus25) yielded better results than min16 and the age group above 18 (plus18). Secondly, more relatable to this thesis, they discovered that when they trained the classifier with four classes, both age, and gender, balanced the dataset based on these classes, reduced the categories to two age classes, and performed the binary classification, they achieved the best results. This implies that age and gender classification could be improved by introducing gender when classifying age and possibly the other way around. The best results were achieved with a balanced dataset with regards to age and gender, with an accuracy of 0.888 and an F_1 score of 0.917 for the adults. Thirdly they conclude that given only 50% of the dataset, the difference in performance was negligible, but for 10% the performance was affected more significantly, still providing better results than a coin flip. They explain how other works in this area have put a heavy emphasis on the lengths of the texts required by each author to get a reliable classification. (Luyckx & Daelemans, 2010) found a drastic drop in scores with regards to the performance of their classifier when the words per text decreased below a hundred. Other works, such as (Burrows, 2007) and (Sanderson & Guenter, 2006), imply that one needs as a minimum requirement, 10000 or 5000 words respectively per author to be able to train the classifier when classifying into specific authors.

2.2.2. Ensemble

(Fauzi & Bours, 2020) used an ensemble method for sexual predator identification in online chats. In their work, they applied various algorithms and feature extraction methods and compared how well they performed on the PAN12 dataset. (Fauzi & Bours, 2020) discovered that an ensemble method combined with a two-stage classifier while using soft voting in the first stage and Naïve Bayes (NB) in the second stage, yielded results that would have granted the first place in the PAN12 competition with an F_1 score of 0.9348. They, in turn, based their study on the research done by (Kittler, 2002), (Larkley & Croft, 1996) and an earlier study by (Fauzi, 2018), which led to the ensemble method with soft and hard voting, to improve the performance of the classifiers. This is a concept that this thesis will explore in the coming chapters, not as in ensemble of different methods, but rather hard and soft voting for a classifier to reach a decision. An example of how to use the classifiers in the way they are used today, applied to four classes, and aggregate the results in attempt to enhance the performance through refining further and working with the different results themselves. What also differs in their research as opposed to this thesis, is that although they used an ensemble method and the classifiers implemented here, they did not look at age and gender as the two-part classifier, but rather predator/non-predator-chat detection and victim/predator within those chats labeled as predatory, similar to (Villatoro-Tello, et al., 2012).

(Kowsari, et al., 2020) is one of the most recent works within the field of gender detection using small texts from one of the most popular social media platforms, Twitter. The dataset was not well balanced in the different classes, which forced an introduction of The Matthews Correlation Coefficient (MCC), to balance the results. As with all methods made for balancing datasets and results, it introduces potential pitfalls or errors, not necessarily mentioned in the paper. This is different from the dataset used in this thesis, as the number of fake accounts, or several accounts per user has skyrocketed in later years. Twitter, as with almost any social platform, is based on a varying amount of trust with regards to both what the user posts, with a certain degree of censoring, and what the user says in its profile. As mentioned earlier, this is akin to one of the objectives of this thesis, to be able to verify, or at least to a degree determine, if the information given by the user is true.

(Kowsari, et al., 2020) used a method based on the Random Multimodel Deep Learning (RMDL) method, which is applicable for many data classification tasks, and in this instance used for text and document categorization. They also made use of different methods for feature extraction and ensemble deep learning for training the model. To do the weighting they used the Adam optimizer, which is known for being computationally efficient, especially when processing large data sets. It is simple to implement and does not require costly hardware, Central Processing Unit- (CPU) or memory-wise.

The results they achieved with RMDL was an F_1 score of 0.8583. (Kowsari, et al., 2020) do not compare their results by the other numbers shown in this chapter, but instead only state that they improve on the results in the field.

The notion of Big Data has been around for some time. It is not mentioned by name in recent works regarding gender and age detection, but it seems as it is still one of the main concerns when it comes to this topic. The growing amount of data found, and given, in all social media platforms, and online for that matter, directly affects both the problem that is malicious users, anonymity, the safety of users and the detection of unwanted behavior. The large amount of data requires effective, automatic, accurate, and reliable detection, that can be used on e.g., social media platforms, not necessarily calling for a Google server park just for this reason alone. There is an immense number of tweets, posts, pictures, snaps, and Tik Toks posted every second of every day, and people tend to not always take personal safety into account.

2.2.3. A broader view of the field

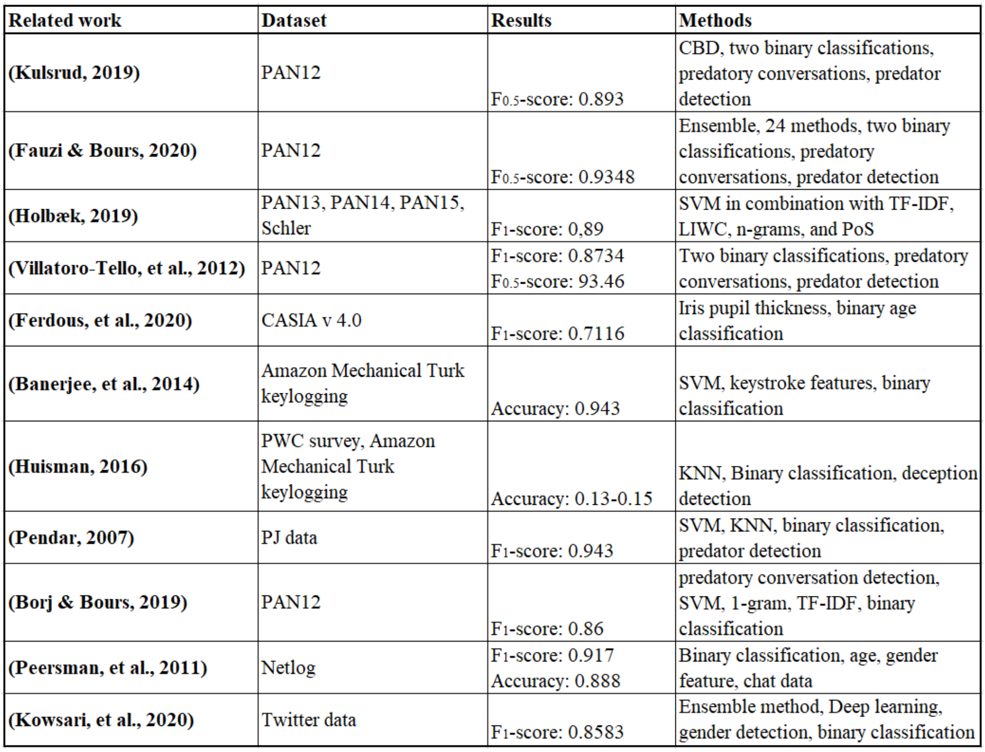
The research regarding the classification of age and gender using machine learning stretches far beyond the use of text and language. (Ferdous, et al., 2020) imply, in their research on age classification using Iris-Pupil thickness, that binary classifications are best suited for age and gender classification, achieving an F_1 score of 0.7116 with the CASIA version 4.0 dataset. Their research is not directly applicable to this study as they used physical biometrical data, but the methods were similar, and they based their work on research done in e.g., age determination using machine learning in social networking and forensics.

As mentioned earlier, not all research that could help in the area of predator detection is specifically designed for this purpose. (Stoll, et al., 2020) focused on detecting impoliteness and incivility in online discussions. The research is done on a dataset consisting of comment sections from German media outlets on Facebook and seeks to detect incivility, covert offensive behavior rather than bold outright offensive comments. (Kalch & Naab, 2017) and (Papacharissi, 2004) claim that behavior such as this, subtle offensive comments, could affect the reader even more than the blunt kind of comments. This is not within the scope of this thesis but would be very interesting to look at from a psychological angle, how much the victims of predators are affected from their experience if the predator were to indirectly groom and approach the victim rather than outright trying to take advantage of the victim. This could change the focus of this field and what kind of behavior and text should be detected.

(Stoll, et al., 2020) also claims that research based on the English language alone is not straightforwardly applicable to other languages such as German, or Norwegian for that matter. The methods they based their work on consisted of several methods and other research. They used their feature sets to create models that focused on finding incivility and impoliteness, using unigrams and n-grams, looking at the words without their meaning, and a lexical approach by tagging specific words as e.g., insults or polarity by using different dictionaries. They also incorporated Named-Entity Recognition and NLP to compare how well the different techniques performed. Their study showed that concepts such as impoliteness and incivility are subjective rather than objective, like lawbreaking behavior, which led to their research producing results with a high misclassification rate. Their best results can be divided into two parts, one part for impoliteness and the other part for incivility. While both parts suffered from the poor results overall due to the subjective matter that is how people use and understand their language, the research showed that the methods used later in this thesis, to some degree, also worked best in determining if a word was impolite or uncivilized. The results favored BoW unigrams and NB, which emphasizes the finding that complex topics such as grooming, impoliteness, bullying, wooing, and manipulation are very hard to detect and shows why the field has been working on building dictionaries for such topics. The best results they achieved were two-parted, an F_1 score of 0.85 for incivility and an F_1 score of 0.66 for impoliteness.

They mention Deep Neural Networks as a method that could improve their research, which would require too much labeled data for them to produce, while also highlighting the need for good datasets, which in their case should have included more complex incivility and impoliteness for their method to detect that sort of behavior better.

Table 1 below shows an overview of the most important related work mentioned in this chapter and the most interesting results achieved by them with regards to this thesis.



Related work	Dataset	Results	Methods
(Kulsrud, 2019)	PAN12	$F_{0.5}$-score: 0.893	CBD, two binary classifications, predatory conversations, predator detection
(Fauzi & Bours, 2020)	PAN12	$F_{0.5}$-score: 0.9348	Ensemble, 24 methods, two binary classifications, predatory conversations, predator detection
(Holbæk, 2019)	PAN13, PAN14, PAN15, Schler	F_1-score: 0,89	SVM in combination with TF-IDF, LIWC, n-grams, and PoS
(Villatoro-Tello, et al., 2012)	PAN12	F_1-score: 0.8734 $F_{0.5}$-score: 93.46	Two binary classifications, predatory conversations, predator detection
(Ferdous, et al., 2020)	CASIA v 4.0	F_1-score: 0.7116	Iris pupil thickness, binary age classification
(Banerjee, et al., 2014)	Amazon Mechanical Turk keylogging	Accuracy: 0.943	SVM, keystroke features, binary classification
(Huisman, 2016)	PWC survey, Amazon Mechanical Turk keylogging	Accuracy: 0.13-0.15	KNN, Binary classification, deception detection
(Pendar, 2007)	PJ data	F_1-score: 0.943	SVM, KNN, binary classification, predator detection
(Borj & Bours, 2019)	PAN12	F_1-score: 0.86	predatory conversation detection, SVM, 1-gram, TF-IDF, binary classification
(Peersman, et al., 2011)	Netlog	F_1-score: 0.917 Accuracy: 0.888	Binary classification, age, gender feature, chat data
(Kowsari, et al., 2020)	Twitter data	F_1-score: 0.8583	Ensemble method, Deep learning, gender detection, binary classification

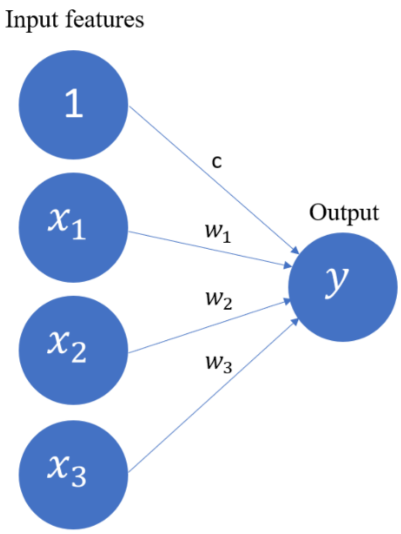
Table 1 an overview of the main related works

2.3. Technical Background

This section will provide the necessary technical background in order to understand the experiment and give the reader an overview of the used machine learning classifiers, preprocessing techniques and types of features used to compare the binary classifiers and the 4-class classifiers.

2.3.1. Logistic Regression (LR)

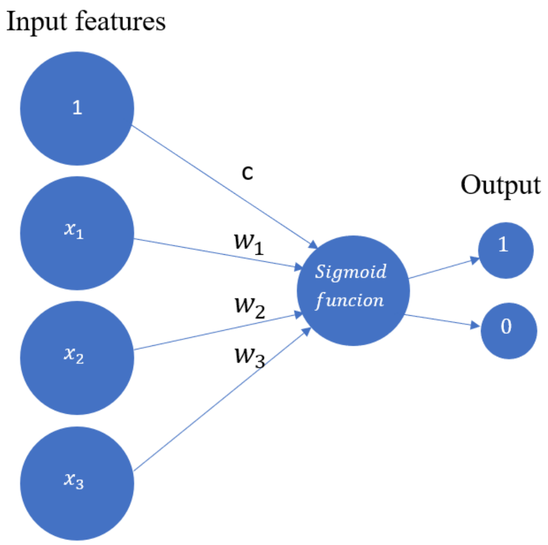
LR is one of the methods categorized as supervised learning within machine learning. It is similar to linear regression in the way that it takes some independent input variables, or predictors as (Navilani, 2019) refers to them, weighs them, and calculates an output, as described in Figure 1 below:



$$y = c + x_1 * w_1 + x_2 * w_2 + x_3 * w_3 + \ \ldots + x_n * w_n$$

Figure 1 Linear Regression, based on (Singh, 2018)

LR, as seen in Figure 2, differs from linear regression as it implements a step in between. It runs the results from the calculated output in linear regression through a sigmoid function, or logistic function. This ensures that the output is between 0 and 1, or -1 and 1.



$$y = \frac{1}{1 + e^{-(c+x_1 * w_1 + x_2 * w_2 + \; x_3 * w_3 + \; \dots \; + x_n * w_n)}}$$

Figure 2 Logistic Regression, based on (Singh, 2018)

LR is normally used for classifications with a binary outcome but can also be used for multinomial classifications. This includes a wide area of possible applications, including classifying the author of sentences into gender and age, both binary and multinomial classifications. Both classifications are being done in this study. (Edgar & Manz, 2017) mention in their book that they use LR for detecting cyber-attacks. As they try to determine if a new sample of a possible attack is indeed the best fit for the "attack" class, the sentences in this study will be determined to be the best fit for both age and gender.

(Subasi, 2020), (Seufert, 2014), and (Gudivada, et al., 2016) all emphasize the applicableness of LR with regards to dichotomous classification, or binary classification problems, e.g., yes/no, true/false, young/old or male/female, and continue to compliment the algorithm for its fast and easy implementation with regards to effectiveness and ability to handle large datasets.

(Navilani, 2019) mentions two possible disadvantages regarding the use of LR. The method could be prone to overfitting and is unable to handle a large set of different features. (Oxford University Press, 2021) describes overfitting as the problem where the results, or analysis, are too closely linked to its dataset, so that the analysis cannot be used or is of little value when introducing other datasets.

2.3.2. Naïve Bayes (NB)

NB classifier is based on Bayes Theorem. (Misra & Li, 2020), who used this method to characterize fractures by classifying sonic waves sent and reflected by different fractures, chose this method because of its simplicity, ability to handle large datasets with high dimensionality, and processing speed.

(Mushtaq & Mellouk, 2017) and (Misra & Li, 2020) both chose NB for its processing speed, which they attribute to what is referred to as the *naive* part of NB, namely its assumption that each feature in the dataset has a conditionally independent contribution to the probability of the classification of a sample. This makes the computations simpler and faster. The classifier computes the conditional probability shown in Equation 1:

$$p(C_k|\ x_1, x_2, \dots, x_n)$$

Equation 1 Conditional probability

Where C_k is the class, *k* specifies which class, and *x* represent the features. The final formula as shown in Equation 2:

$$\begin{aligned}
&p(C_k|\ x_1, x_2, \dots, x_n) \\
&\propto\ p(C_k|\ x_1, x_2, \dots, x_n) \\
&= p(x_1|C_k) \cdot p(x_2|C_k) \dots \cdot p(x_n|C_k) \cdot p(C_k) \\
&= p(C_k) \prod_{i=1}^{n} p(x_i|C_k)
\end{aligned}$$

Equation 2 final formula for Naïve Bayes Classifier

Shows where the Bayesian theorem has been applied under the assumption that all the features *x* is mutually independent. (Singh, et al., 2019) explains the difference between multinomial NB and Bernoulli NB as the former considers the feature vector where the terms represent the frequency of which it appears, while the latter only considers the feature in a binary fashion, if the term appears or not.

2.3.3. Neural Network (NN)

NNs were originally inspired by the human brain and how it functions. (Marini, 2009) explains that there are two paths of science within NN, one which focuses on mimicking and understanding the human brain, and one more focused on computations. The computational NNs have proven themselves as able to solve and compute difficult problems not easily solved, or currently impossible to solve otherwise with traditional mathematics and statistics. NNs are being used in areas such as predicting the weather, signal filtering and in this case: classification of certain patterns.

A simple representation of a NN, or Neural Classifier, normally consists of an amount of input nodes, a hidden layer of nodes, and an output layer or node, as can be seen in Figure 3 Figure 3 A simple representation of a neural network below. The features, or variables, are introduced in the input nodes, which are then forwarded to the hidden layer. The hidden layer nodes apply a non-linear transfer function to the sum, which has been differently weighted for each node, and forwards it to the output layer or node. The output layer again receives a weighted sum from the hidden layer and applies another non-linear transfer function. The output could be a number between 0 and 1, which for a binary classification could be if the input belongs to one class or the other.

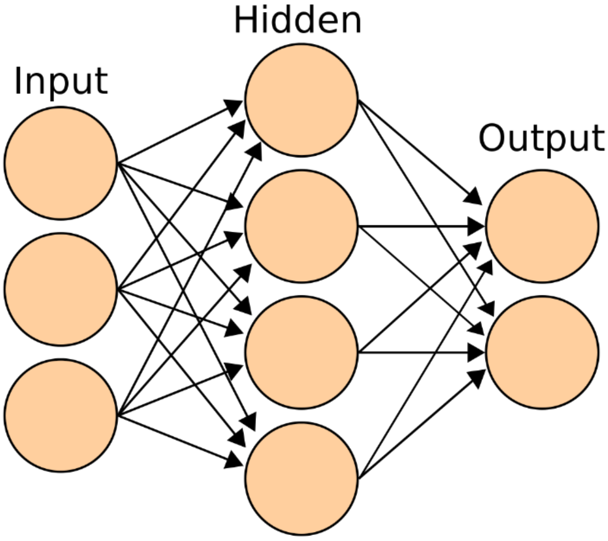


Figure 3 A simple representation of a neural network (Burnett, 2006)

NN does have some drawbacks compared to the other methods in this study. (Miner, et al., 2012) and (Bunge & Judson, 2005) mention some of them as NN being computationally heavy, as the algorithm and computations are considerably more complex, which in turn makes the method or analysis of how the results were achieved difficult to understand. NNs are often referred to as a “black box”, and a significant amount of research have been undertaken to improve processing speed of computers to be able to cope with the computation time required for advanced NNs.

2.3.4. Decision Tree (DT)

A DT is a predictive model that consists of three types of nodes, often called the root node, leaf nodes, and split nodes. As Figure 4 illustrates, the root and split nodes decide one of two outcomes, while the leaf nodes represent the decision made by the tree, here being whether to go on a hike or not. DTs can be used for multi-class classification as well, by e.g., assigning each class a specific integer and several threshold values instead of using a binary decision where we have either yes or no where we only need one threshold value.

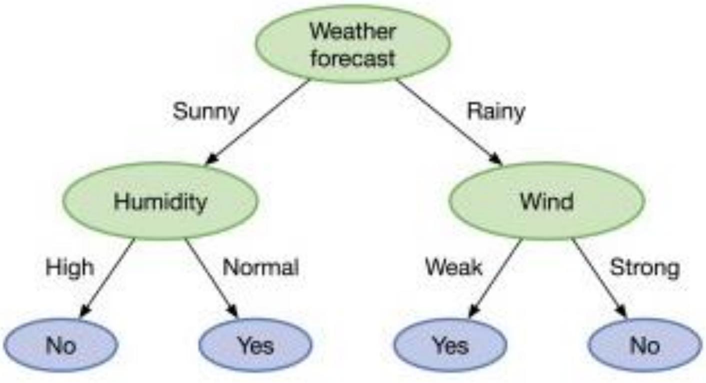


Figure 4 A simple decision tree (Reinders, et al., 2019)

(Kotu & Deshpande, 2015) provide a list of what they present as distinct advantages of using DTs. It includes ease of interpretation, hardly any data preparation, feature selection is done by the tree itself implicitly, and the performance of the tree is not affected by possible nonlinear relationships between the data.

(Tan, 2015) explains one of the common disadvantages with DT, and other classification methods, which is overfitting. It is easy to build a DT for a given dataset, which often results in deep trees, with complex decision rules and more fitting to a specific dataset, but more of a challenge to build good DTs with shorter branches. A frequently used method to mitigate overfitting is overfitting pruning. This should make the tree(s) applicable for unlabeled data and implement tolerance for errors from wrongly labeled training data.

2.3.5. Random Forest (RF)

RF, as (Reinders, et al., 2019) and (Gedeck, et al., 2010) explain, consist of several simple DTs, and is one of the methods referred to as ensemble based. While DT comes to a decision within one tree, RF uses several DTs in the training stage and e.g., performs a majority vote with all the decisions from the trees in the classification stage. This is also one of the methods that inherently mitigates the overfitting issue related to DTs, whereas pruning is not applied here. Even though DTs are faster, RF still retains much of the coveted speed. (Dramsch, 2020) commends RF for its ability to become very complex and useful predictive models in geophysics. As Figure 5 below illustrates, the RF takes the decisions of all the trees, and in this case performs an average, which in turn yields a total decision.

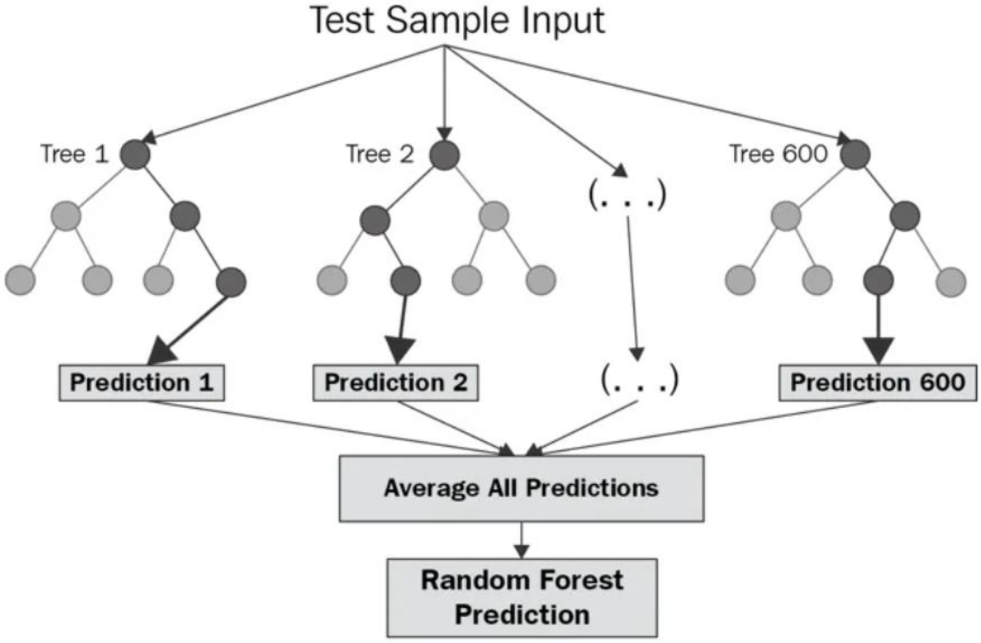


Figure 5 An example of a RF (Chakure, 2019)

2.3.6. Features

2.3.6.1. Bag of Words (BoW)

BoW is a model that extracts features from text, which in this study is for use in text classification. The model needs a dictionary of known words in order to compare the terms, or words, in the text. When it finds a known word it remembers only the occurrence of it, not where or in which context. The model simple and easy to implement, and (Brownlee, 2017) mentions that complexity can be introduced through how the occurrence of words is scored through term weighting methods, such as Binary, Term Frequency (TF), and Term Frequency-Inverse Document Frequency (TF-IDF), or what kind of dictionary is used.

One of the disadvantages of BoW is the exclusion of context. All the words are put in a “bag”, meaning that the order of words in a sentence is discarded. The sentence “Protect children from predators” would in the eyes of BoW be the same as “Protect predators from children”.

2.3.6.2. Term Weighting Methods

This study uses three different term weighting methods, as mentioned above. Binary being the easiest approach, as it only scores if the word appears in the text or not. TF takes note of how many times a word appears in a text, while TF-IDF is the most complicated one but the method that contains the most information as well. TF-IDF tries to give a score to the different terms based on how often they occur in all the texts, or documents, which says something about how rare the word is in the given dataset.

Given the keywords of this thesis:
Machine learning, Age and gender detection, 4-Class Classification, Multi-class classification, 2-Class Classification, Binary Classification, Chatroom security

Binary would represent it as (if all words/terms was in the dictionary):

Machine, learning, Age, and, gender, detection, 4-Class, Classification, Multi-class, 2-class, Binary, Chatroom, Security.

It would not say anything about the order in which they were found, the context, or how many times they occurred.

TF could be represented as a histogram as seen in Figure 6 below, recording how many times the terms occur.

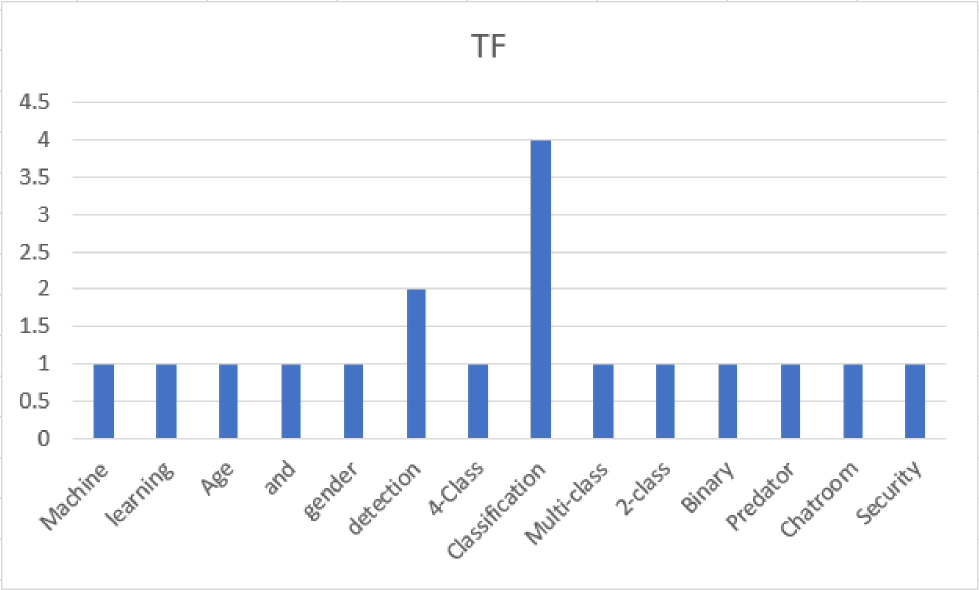


Figure 6 example of a TF representation

TF-IDF can be used for an array of different things. In text classification it can be used to find the most important word of a document (set), which would imply that it is a keyword of that document. This would be the word(s) with the highest TF-IDF score. In predator detection, using a chat written by a predator, those words could be words to look for, used as indicators in other chats to detect the presence of a similar predator. Another use case could be search engines, where the search word is used to show relevant documents or results based on the TF-IDF score for that specific word.

3. Methodology

This chapter provides an overview of the chosen methodology for this study, how the literature was selected, and how the data was collected and processed.

3.1. Literature study

A literature study was conducted in order to gain a required level of knowledge of the current state of the art. Some research was specifically selected as this study builds on its ideas and results, while others were selected based on specific parts of the research that overlapped with this study, either with regards to methods used, results achieved, the dataset used, or to provide a broader view of the field. The main body of literature was found in scientific databases, books, articles, or theses.

3.2. Data collection

The dataset chosen for this study is the Schler dataset. This is both to ensure comparability with related works and because it is one of the main datasets used for text classification. Collecting and processing an original dataset would be very time consuming, and factors that is guaranteed to affect the results would be e.g., age groups available, language differences between the new dataset and the ones used by others, especially vocabularies and of course the platforms where the data is gathered from. By using the Schler dataset, which consists of blogs, that is used by other research for the very same purpose, those factors will not have an impact on the results, which will make them more comparable. One possible disadvantage could be that the data can be somewhat outdated, as the differences in language, and vocabularies in conjunction with age groups varies over time.

Other datasets could've been chosen, such as the PAN datasets mentioned in Table 1, but as this study is more focused on age and gender detection rather than predator detection, the research more likely to have comparable results, would be the ones using e.g., the Schler dataset.

3.3. Data analysis

The first of the two main tasks for the data analysis part, regarding extracting the features and training the different models on the training data and applying this on the testing data, was done by co-supervisor Mohammad Ali Fauzi. The second main task, the processing, final classification, and analysis of the results from the 18 different methods, consisting of algorithms and feature extraction methods will solely be produced by this study. By not running the testing and training as part of this thesis, it ensures that the initial classification method is equivalent to that of (Fauzi & Bours, 2020), which is an article with very good results. It also removes one of the possible pitfalls by not introducing new sources of error by using other coding and scripts. Writing the necessary code and scripts for the testing and training is not within the scope of this thesis, but rather the code and scripts for processing those results.

The results produced and achieved by the methods mentioned above will be processed in several ways. The main task for this thesis is to compare how well the 4-class classifiers holds up against the two binary classifiers. In order to make the data comparable it must, in some way, become standardized. Hence, confusion matrices will be established as a first attempt, and overviews in form of graphs and tables will be constructed. The data from the two binary classifiers only has two possible outcomes, which makes the probability less complex to deal with, where the age classification is the probability of it being a child, while it is the probability of it being a man for the gender classification. The 4-class classifier produces four outputs, where each class, different from with the binary classifiers, is given a specific probability, meaning that each class; male child, male adult, female child, and female adult, has a score between 0 and 1. For this to make sense, the comparison of the gender class will be done by the addition of the two gender classes in the four-class classifier, then comparing this with the binary class gender classification. The equivalent will be done for the age classes. Another way the data will be processed is by soft and hard voting. If the scores per sentence is soft voted, producing one score per author, the result is hard voted, or rounded to give a final classification. If the scores per sentence is hard voted, then an average is calculated, and a rounding, or hard vote, of that result is performed to give a final classification. Figure 7 below shows how it will be done for the binary sentence classification scores.

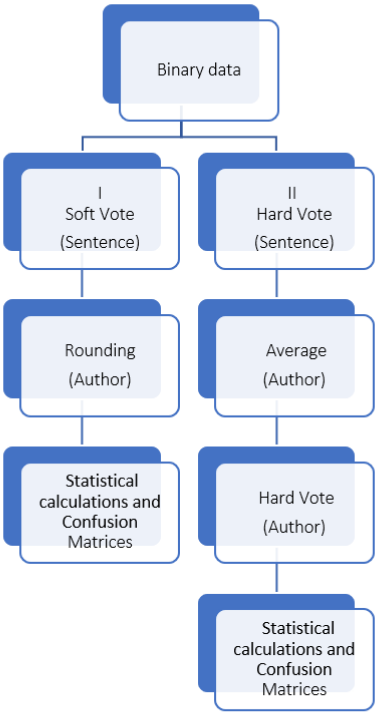


Figure 7 An overview of the processing methods applied to the binary classification dataset

The binary data consisted of all the methods applied for binary classification on the given dataset. Soft voting implies that an average of all the scores for the different sentences has been taken per author. Rounding after means that if the average was lower than 0.5, or 0.5 and above, it was set to 0 or 1 respectively. On the other hand, hard voting on the given dataset sets the class of the sentence to 0 or 1 based on the same limits, and the average of all the sentences per author is performed with those results. The hard vote at the end is just for rounding the numbers and setting the final class for each author, either rounding up from 0.5 and above, or rounding down from scores lower than 0.5.

The process is approximately the same for 4-class combined data classification, as the same process is applied after the combination of the date, but different for the 4-class data classification. The difference for 4-class classification is that the initial hard vote sets the highest scoring class to 1 and the rest to 0. The same is applied after soft voting is done initially and the hard vote is perfomed to classify the author.

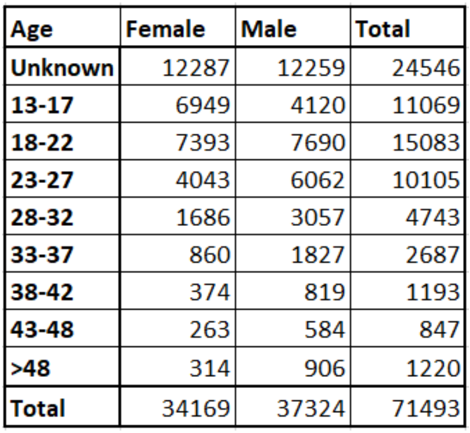
The data preparation, parsing, and analysis for this stage will be carried out through custom-written scripts in R and Bash, while the computation of accuracy, precision, recall and F scores will be done in Microsoft Excel.

4. Data description

(Schler, et al., 2006) created one of the popular datasets used for text classifications in 2004. The main takeaway from their research was the differences between the writing styles of both age and genders, based on the vocabularies used by the authors. This is directly applicable to this thesis, which is why this dataset should be able to provide a good basis to answer the research question.

The dataset was gathered from blogger.com in August 2004. They gathered all the blogs available throughout one day, removing all who did not have at least 200 recognizable English words, and those who did not have an indication of the gender of the author, giving a total of over 71000 blogs. They mention that some of the blogs had an indication of age as well.

As can be seen from Table 2 below, the distribution of blogs with regards to gender and age is not balanced. Therefore, a sub corpus of this has been extracted for this thesis as to have the same amount of male and female during the first stage where the probabilities are calculated. This was also done by (Schler, et al., 2006) which left them with over 37000 blogs with over 295 million words.



Age	Female	Male	Total
Unknown	12287	12259	24546
13-17	6949	4120	11069
18-22	7393	7690	15083
23-27	4043	6062	10105
28-32	1686	3057	4743
33-37	860	1827	2687
38-42	374	819	1193
43-48	263	584	847
>48	314	906	1220
Total	34169	37324	71493

Table 2 Author distribution in the Schler dataset

It is noteworthy that these blogs are gathered from 2004, which could influence the classification results, potentially making them a bit less applicable in today's chats. Also, chats might be generally less formal than blogs. Blogs is furthermore not a dialogue, but rather a monologue, which means that it is probably structured a bit differently, even if this study looks at every sentence by itself. Nevertheless, the answers to the research questions in this study should be unaffected by this, as both the binary and 4-class classifiers are equally affected by using the same data and should be applicable to a more present-day dataset as well.

While not directly applicable to this study, (Schler, et al., 2006) show some of the differences between the various ages and genders. They continue to draw conclusions about why they occur, relating some of the differences to how a normal life of a person with resources enough to maintain a blog transpires. The focus of a student is not the same as that of people in later stages of life, which correlates more to work, finance, and marriage. One of the specific differences between the genders is the use of articles and prepositions, which are more used by male authors, and assent or negation words and pronouns, which are used more by female authors.

The dataset used in this study is the results from Muhammad Ali Fauzi’s work on the Schler dataset. Fauzi performed preprocessing, feature extraction, model training and testing.

Normally there are several steps included in preprocessing. (Uysal & Gunal, 2014) mention tokenization, stop-word removal, lower or uppercase conversion, and stemming as the most common steps. Only tokenization was performed on this dataset, splitting the blogs into tokens, or terms. A term could be a word, number, an emoji or alike.

The feature extraction for this study made use of BoW, as explained in the technical background, meaning that all unique words, or terms in the training data, are used as features without looking at the order in which they appear.

The Vector Space Model (VSM) was applied on the blogs with the BoW features. (Holle, et al., 2015) explains it as VSM representing the blogs as feature vectors in space terms. Every word, now represented as a feature, has a value calculated through three different term weighting methods. These methods are Binary, TF, and TF-IDF:

1. Binary

   Every blog *c* was represented as a binary vector $\underset{v}{\rightarrow} = \left(b_{1,c}, b_{2,c}, \dots, b_{N,c}\right)$ where $b_{t,c} = 1$ if a term *t* appeared in a blog, and 0 if it did not.
2. TF

   For TF a blog was represented as a count vector, meaning that a blog *c* was represented by the vector $\underset{v}{\rightarrow} = \left(tf_{1,c}, tf_{2,c}, \dots, tf_{N,c}\right)$ where $tf_{t,c}$ counts how many times a term appears in the blog.
3. TF-IDF

   Every blog *c* was represented by a $tf \cdot idf$ vector $\underset{v}{\rightarrow} = \left(tf \cdot idf_{1,c}, tf \cdot idf_{2,c}, \dots, tf \cdot idf_{N,c}\right)$ where $tf \cdot idf_{t,c}$ represents the weight of term *t* in blog *c*. Fauzi used normalized TF and IDF. This means that if terms t_1 and t_2 in blog *c* occurs 1 and 3 times, having TF-values of 1 and 3, indicating that t_2 is more important in blog *c* than t_1, it does not mean that t_2 is three times more important than t_1. Fauzi normalized the TF using Equation 3 normalized TF below:

$$wtf_{t,c} = \begin{cases} 1 + \log tf_{t,c} & tf_{t,c} > 0 \\ 0 & otherwise \end{cases}$$

Equation 3 normalized TF (wtf)

The IDF, which represents the rarity of the given term, gives a higher score, the less the term is used in the blogs. Equation 4 IDF below shows the computation of the IDF, where *N* is the number of blogs, and df_t is the number of blogs the term *t* appears in:

$$idf_t = \log \frac{N}{df_t}$$

Equation 4 IDF

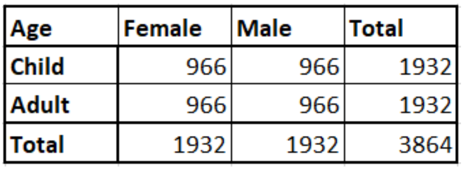
By using the normalized TF and IDF depicted, the TF-IDF value for term *t* will increase with the number of occurrences in a blog, and the rarity within the dataset, as Equation 5 TF-IDF shows:

$$tf \cdot idf_{t,c} = \ wtf_{t,c} \cdot idf_t$$

Equation 5 TF-IDF

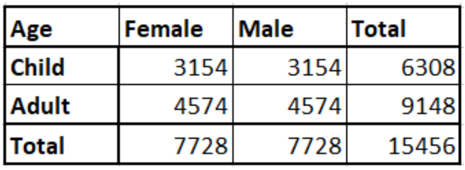
As mentioned, the classification used three feature sets consisting of Binary, TF and TF-IDF features. There were used six machine learning algorithms: Multinomial NB, Bernoulli NB, DT, LR, NN and RF, resulting in 18 different methods, or combinations. The data was split into sentences instead of entire blogs, where every sentence was linked to its author. The data was furthermore split into a test and a training set. The classification was done in three ways, binary age classification, binary gender classification and 4-class classification. So, for every combination of classification, type of feature set and algorithm, a model was trained on the sentences from the training data, and testing was performed with the testing data and provided to this study. This means that for the binary classification there are 18 scores per sentence, and 72 scores per sentence for the 4-class classification, as it has four scores, one for each class per combination. This study seeks to answer the research question based on these scores.

When training the model, Fauzi first extracted a training set. This set consisted of 3864 people who wrote exactly 1895108 sentences together. There was no overlap between the training and the testing dataset, but there are differences in the number of sentences per author for both sets. The author distribution for both the training and the testing set used by Fauzi, and by extension this study, can be seen below in Table 3 and Table 4:



Age	Female	Male	Total
Child	966	966	1932
Adult	966	966	1932
Total	1932	1932	3864

Table 3 distribution of authors in the training dataset

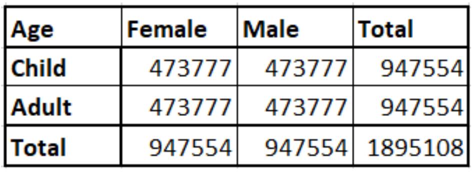


Age	Female	Male	Total
Child	3154	3154	6308
Adult	4574	4574	9148
Total	7728	7728	15456

Table 4 distribution of authors in the test dataset

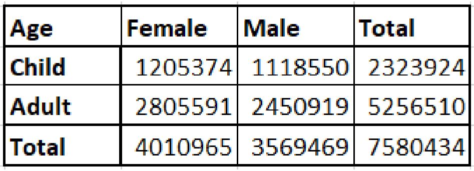
Where child represents ages from 13 to 17 and adult represents both ages 23-27 and 33-42.

As mentioned, the number of sentences per author varies. The distribution of sentences with regards to age and gender, for both the training and test dataset, is as Table 5 and Table 6 shows:



Age	Female	Male	Total
Child	473777	473777	947554
Adult	473777	473777	947554
Total	947554	947554	1895108

Table 5 distribution of sentences in the training dataset



Age	Female	Male	Total
Child	1205374	1118550	2323924
Adult	2805591	2450919	5256510
Total	4010965	3569469	7580434

Table 6 distribution of sentences in the test dataset

5. Data analysis and results

This chapter presents and discusses the results from the methods explained in chapter 3. Due to the large number of results, only a subset of the results will be presented in this chapter. Appendix A will show all the created tables with the obtained results. The results are firstly presented individually for the binary classifiers, then the 4-class classifiers, and lastly compared with each other at the end of the chapter. Some confusion matrices have been included, while the rest is shown in Appendix A.

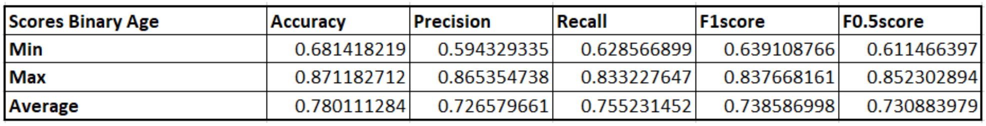
5.1. Results of the binary classifiers

5.1.1. Age classifier

The age classifiers include all the different methods used, where three different feature extraction methods, six machine learning algorithms, and soft and hard voting are implemented. For the both the binary and 4-class classifier, the age was determined as the author being an Adult or a Child, where the adult corresponds to the authors in their thirties (33-42) and twenties (23-27) and the children in their teens (13-17) in the dataset. An overview of the different metrics for each method will be shown after both Age and Gender classifiers have been accounted for using the different voting methods. Figure 7 already presented in section 3.3 showed the process followed below.

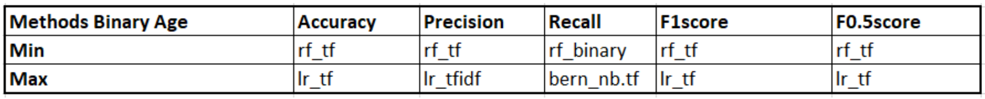
5.1.1.1. Method I (soft vote)

The results from processing method I (soft vote) can be represented as Table 7 and Table 8 below, along with some of the corresponding confusion matrices: Table 9 and Table 10. Appendix table 21 show all metrics and methods for binary age classification soft vote, with highlighting of the highest and lowest scores for each metric. Similar tables can be found in the appendix for all methods with binary classification, both for the binary classification data and the 4-class combined classification data.



Scores Binary Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.681418219	0.594329335	0.628566899	0.639108766	0.611466397
Max	0.871182712	0.865354738	0.833227647	0.837668161	0.852302894
Average	0.780111284	0.726579661	0.755231452	0.738586998	0.730883979

Table 7 The best and worst results from binary age classification using soft voting



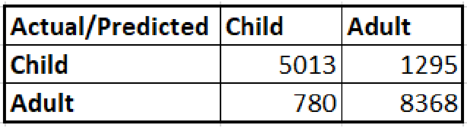
Methods Binary Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tf	rf_tf	rf_binary	rf_tf	rf_tf
Max	lr_tf	lr_tfidf	bern_nb.tf	lr_tf	lr_tf

Table 8 The methods corresponding to the best and worst results from binary age classification using soft voting

Regarding the probability score in the binary methods, 0 corresponds to an adult, and 1 corresponds to a child. Precision could be more interesting, especially in the context of detecting pedophiles. It could be argued that misclassifying a child as an adult (False Negative) is not as bad as misclassifying an adult as a child (False Positive), in that context $F_{0.5}$ score is a more interesting metric than F_1 score, giving more weight to precision than recall.

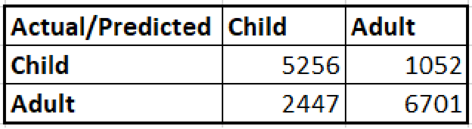
It does seem that RF, especially paired with TF, yields poor results when it comes to the age classification with binary classes using soft voting on the sentences, and LR performs the best overall, while Bernoulli NB yields the best recall.

It becomes apparent that LR yields the best results, depicted in Table 9, both from the objective view of the scores including accuracy, precision, and F scores, and the problem from the contextual view, correctly classifying adults as adults. Even though Bernoulli NB yielded a high recall, it classified nearly 27% of the adults as children, as shown in Table 10.



Actual/Predicted	Child	Adult
Child	5013	1295
Adult	780	8368

Table 9 LR TF-IDF on Binary age data soft vote

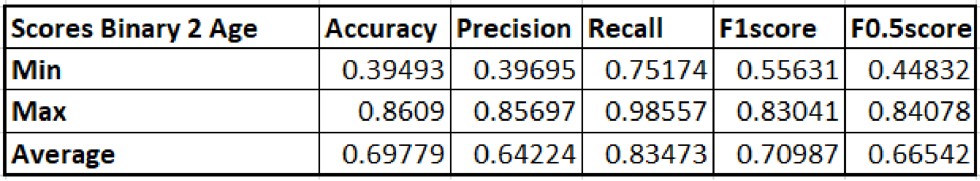


Actual/Predicted	Child	Adult
Child	5256	1052
Adult	2447	6701

Table 10 Bernoulli NB TF on Binary age data soft vote

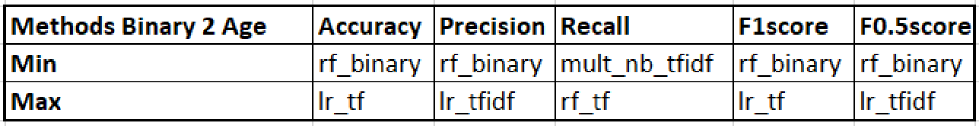
5.1.1.2. Method II (hard vote)

The results from processing method II (hard vote) can be represented as Table 11 and Table 12 below, along with some of the corresponding confusion matrices: Table 13 and Table 14.



Scores Binary 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.39493	0.39695	0.75174	0.55631	0.44832
Max	0.8609	0.85697	0.98557	0.83041	0.84078
Average	0.69779	0.64224	0.83473	0.70987	0.66542

Table 11 The best and worst results from binary age classification using hard voting

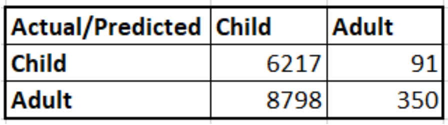


Methods Binary 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_binary	rf_binary	mult_nb_tfidf	rf_binary	rf_binary
Max	lr_tf	lr_tfidf	rf_tf	lr_tf	lr_tfidf

Table 12 The methods corresponding to the best and worst results from binary age classification using hard voting

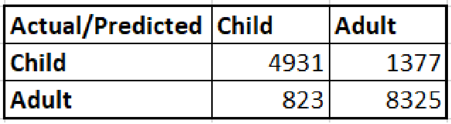
The LR algorithm, with both TF and TF-IDF, as seen in Table 14, yields the overall best results here as well. The poor performance by RF is also repeated here, but mostly for the binary feature set. The best performance with regards to recall is surprisingly RF TF with a score of 0.985, but if we look at the confusion matrix for that method below, Table 13, we see that the method classified 97% of the authors as children, of whom 8798 were adults and 6217 classified correctly as children. The multinomial NB with TF-IDF performed under par with regards to recall but had a decent precision score of 0.837, which as mentioned could be a more important metric.

The LR algorithm with TF-IDF, shown in Table 14, did indeed have the best precision score of 0.856, but it was closely followed by the precision score of LR TF, which was 0.826. Multinomial NB TF-IDF had the worst recall score but managed a precision score of 0.837.



Actual/Predicted	Child	Adult
Child	6217	91
Adult	8798	350

Table 13 RF TF on Binary age data hard vote



Actual/Predicted	Child	Adult
Child	4931	1377
Adult	823	8325

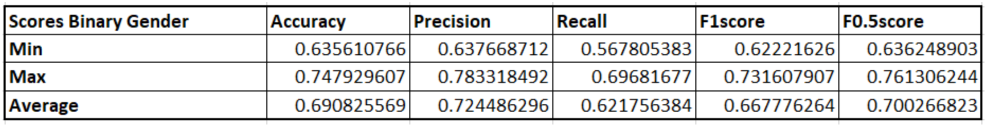
Table 14 LR TF-IDF on Binary age data hard vote

5.1.2. Gender classifier

The gender classifiers use the same approach as described for the age classifiers. The author will be classified as either a female or a male using all the same methods, both with algorithms and with feature methods, including soft and hard voting.

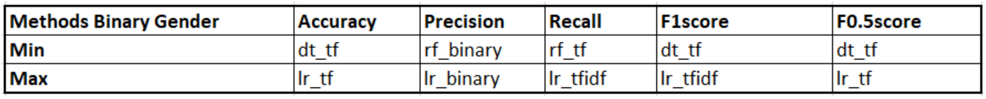
5.1.2.1. Method I (soft vote)

The results from processing method I (soft vote) can be represented as Table 15 and Table 16 below, along with some of the corresponding confusion matrices: Table 17 and Table 18.



Scores Binary Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.635610766	0.637668712	0.567805383	0.62221626	0.636248903
Max	0.747929607	0.783318492	0.69681677	0.731607907	0.761306244
Average	0.690825569	0.724486296	0.621756384	0.667776264	0.700266823

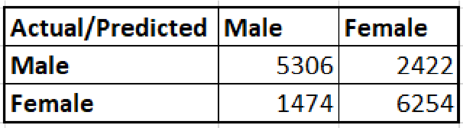
Table 15 The best and worst results from binary gender classification using soft voting



Methods Binary Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	dt_tf	rf_binary	rf_tf	dt_tf	dt_tf
Max	lr_tf	lr_binary	lr_tfidf	lr_tfidf	lr_tf

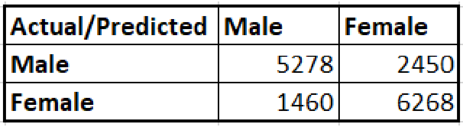
Table 16 The methods corresponding to the best and worst results from binary gender classification using soft voting

The best performing methods here are predominantly LR, yielding a precision of over 0.724 with binary features and an $F_{0.5}$ score of 0.761 with TF features, while DT and RF yield poor results close to 0.6 overall. RF performed similar for the age classifier, but LR performed close to 10% worse for the gender classification compared to the age classification. This indicates that there are more differences between different ages than there are for genders.



Actual/Predicted	Male	Female
Male	5306	2422
Female	1474	6254

Table 17 LR TF on Binary Gender data soft vote

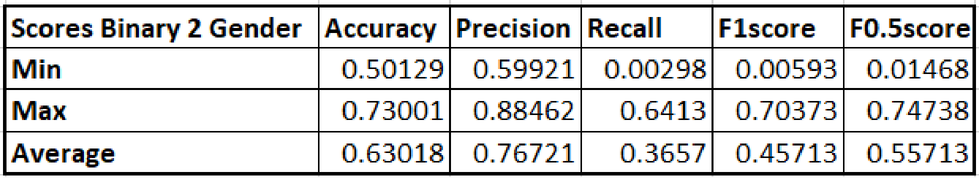


Actual/Predicted	Male	Female
Male	5278	2450
Female	1460	6268

Table 18 LR Binary on Binary Gender data soft vote

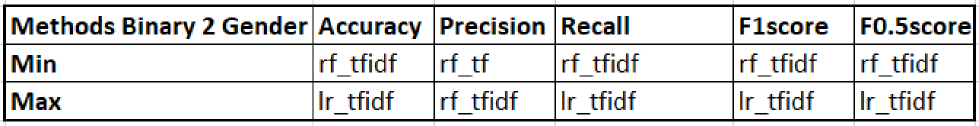
5.1.2.2. Method II (hard vote)

The results from processing method II (hard vote) can be represented as Table 19 and Table 20 below, along with some of the corresponding confusion matrices: Table 21 and Table 22.



Scores Binary 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.50129	0.59921	0.00298	0.00593	0.01468
Max	0.73001	0.88462	0.6413	0.70373	0.74738
Average	0.63018	0.76721	0.3657	0.45713	0.55713

Table 19 The best and worst results from binary gender classification using hard voting

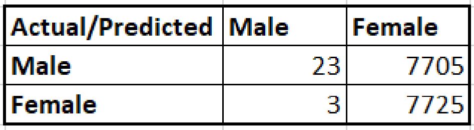


Methods Binary 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tfidf	rf_tf	rf_tfidf	rf_tfidf	rf_tfidf
Max	lr_tfidf	rf_tfidf	lr_tfidf	lr_tfidf	lr_tfidf

Table 20 The methods corresponding to the best and worst results from binary gender classification using hard voting

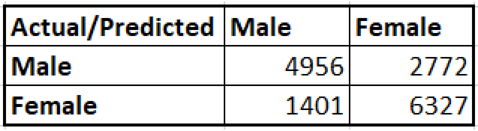
Using hard voting LR still remains the method with the best results, as Table 22 presents, while RF shows the lowest scores for all chosen metrics. RF has close to zero recall, but a high precision, the reason being that almost every author was classified as female, as Table 21 shows.

The scores achieved by the gender classifiers are on average lower than those of the age classifiers, especially the accuracy, recall, and F_1 and $F_{0.5}$ score. While precision score is potentially more important when classifying age, it is not as important for gender, even though that score measure slightly better against the age classification score using hard voting.



Actual/Predicted	Male	Female
Male	23	7705
Female	3	7725

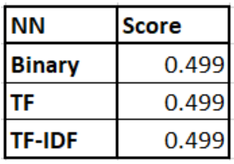
Table 21 RF TF-IDF on Binary Gender data hard vote



Actual/Predicted	Male	Female
Male	4956	2772
Female	1401	6327

Table 22 LR TF-IDF on Binary Gender data hard vote

Neural networks had some issues with the binary classification, as every sentence of every author, regardless of age or gender yielded the results shown in Table 23 before any soft or hard voting:



NN	Score
Binary	0.499
TF	0.499
TF-IDF	0.499

Table 23 scores for NN binary classification

These scores will classify every sentence of every author as either an adult for age classification, or a female in gender classification. The problem did not occur with the 4-class classification data.

The distribution of $F_{0.5}$ scores for both the age and gender classifiers are as shown below in Figure 8 and Figure 9. F scores for NN cannot be calculated as every author is classified either as an adult or a female. The age classification is better than the gender classification in 8 out of 15 methods using soft voting (SV), disregarding NN. The differences between age and gender classification become more apparent for Bern NB, DT and RF when hard voting (HV) is implemented on the sentences initially, while the opposite is true for the other methods.

Curiously the $F_{0.5}$ results for hard voting are the inverse of the precision scores using hard voting, in the sense that the age classification has better results in 11 out of 15 methods, whereas for the precision scores the gender classification has better results in 11 out of 15 methods.

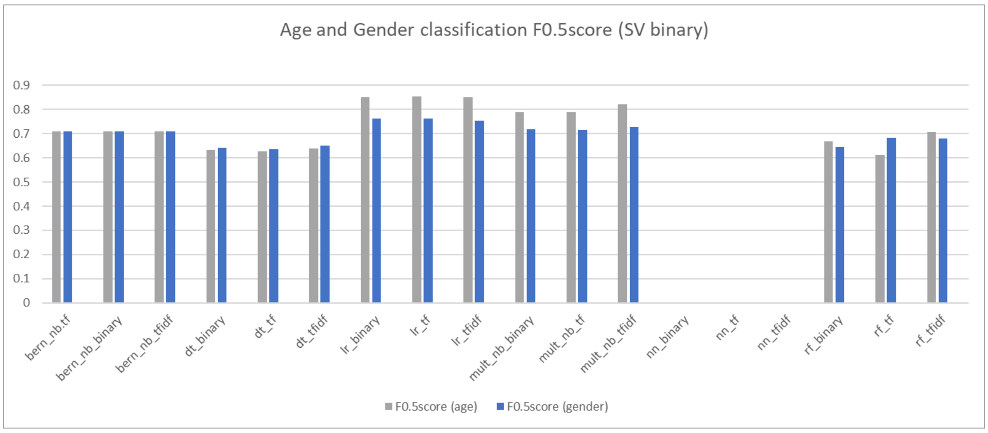


Figure 8 F0.5 scores for Age and Gender classification Binary input data soft vote

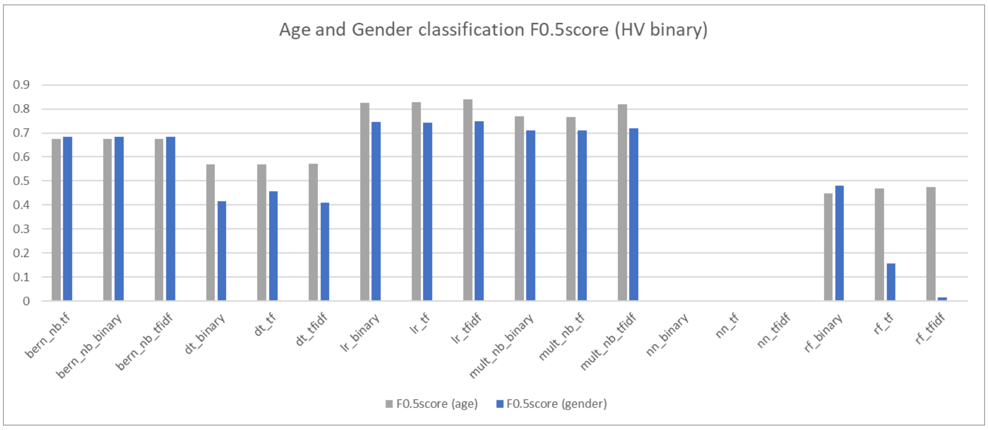


Figure 9 F0.5 scores for Age and Gender classification Binary input data hard vote

5.2. Results of the 4-class classifiers

The 4-class data has been processed in the same way as the binary data as shown in Figure 10 below:

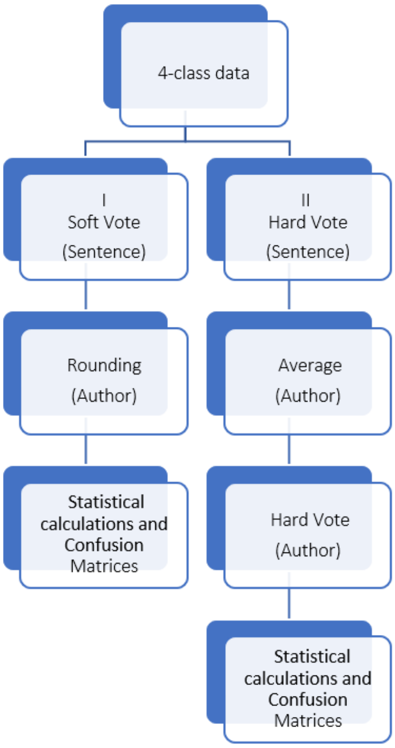
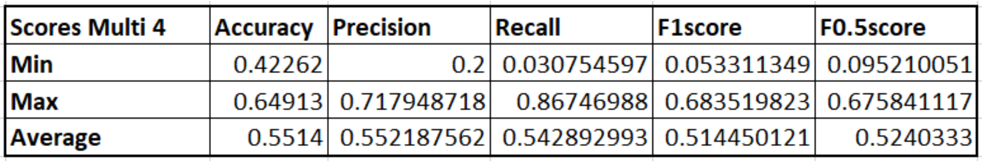


Figure 10 An overview of the processing methods applied to the 4-class dataset

In addition to this, the data has been combined into a binary dataset in order to be able to compare the results with the results from the binary dataset. These results will be presented in section 5.3.

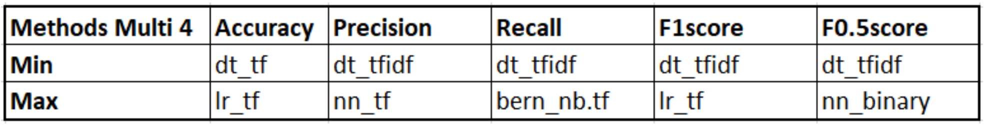
5.2.1. Method I (soft vote)

The results from processing method I can be represented as Table 24 and Table 25 below, along with some of the corresponding confusion matrices: Table 26 and Table 27.



Scores Multi 4	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.42262	0.2	0.030754597	0.053311349	0.095210051
Max	0.64913	0.717948718	0.86746988	0.683519823	0.675841117
Average	0.5514	0.552187562	0.542892993	0.514450121	0.5240333

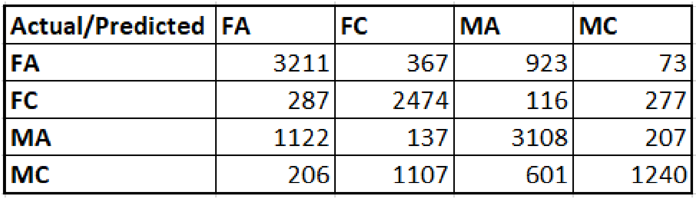
Table 24 The best and worst results from 4-class classification using soft voting



Methods Multi 4	Accuracy	Precision	Recall	F1score	F0.5score
Min	dt_tf	dt_tfidf	dt_tfidf	dt_tfidf	dt_tfidf
Max	lr_tf	nn_tf	bern_nb.tf	lr_tf	nn_binary

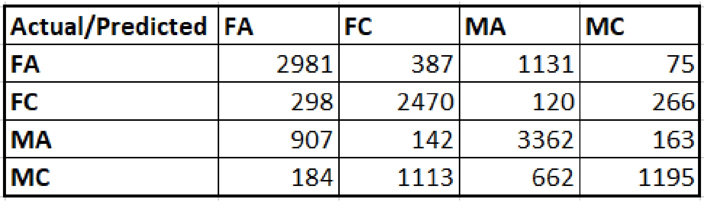
Table 25 the methods corresponding to the best and worst results from 4-class classification using soft voting

It becomes apparent that DT underperforms as it has the lowest scores for all metrics used here. NN performs well, as shown in Table 27, especially with regards to precision and $F_{0.5}$ score, but these can't be compared to the results from the NN binary classification. They can however be compared to the results from other classifiers. Bernoulli NB TF still yields the best results with regards to recall, and LR TF continues to perform well, as shown in Table 26. The accuracy here represents the score based on getting both the age and gender right, which is a difficult challenge. Table 26 and Table 27 below presents the confusion matrices for selected methods based on the tables above. The classes used are: Female Adult (FA), Female Child (FC), Male Adult (MA) and Male Child (MC). There are more links between these classes, as one of the characteristics, either age or gender, always is associated with another class.



Actual/Predicted	FA	FC	MA	MC
FA	3211	367	923	73
FC	287	2474	116	277
MA	1122	137	3108	207
MC	206	1107	601	1240

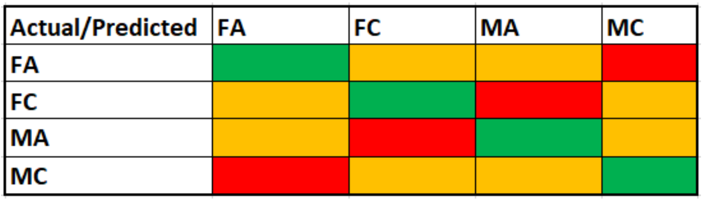
Table 26 LR TF - 4-class classification soft voting



Actual/Predicted	FA	FC	MA	MC
FA	2981	387	1131	75
FC	298	2470	120	266
MA	907	142	3362	163
MC	184	1113	662	1195

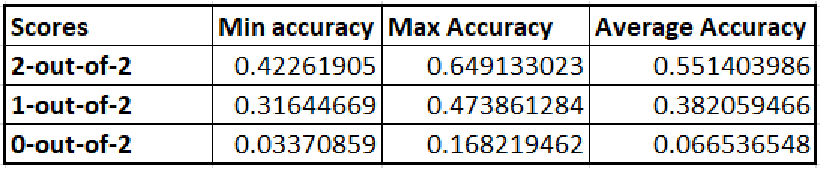
Table 27 NN Binary - 4-class classification soft voting

Table 28 shows how the different are accuracies calculated, where green represents both age and gender classified correctly, red represents the classification where none of the characteristics are correct, and orange where one of two characteristics are classified correctly. Table 29-Table 32 show the results from the different accuracies, based on Table 28. When considering the cases where at least one part of the classification is correct, meaning 2-out-of-2 and 1-out-of-2, NN-Binary comes out on top, with a combined accuracy of almost 97%.



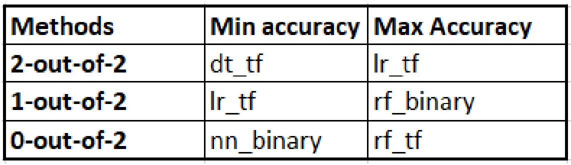
Actual/Predicted	FA	FC	MA	MC
FA	🟩	🟧	🟧	🟥
FC	🟧	🟩	🟥	🟧
MA	🟧	🟥	🟩	🟧
MC	🟥	🟧	🟧	🟩

Table 28 Accuracy legend



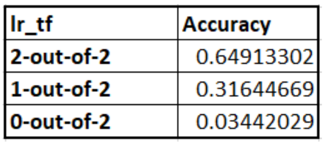
Scores	Min accuracy	Max Accuracy	Average Accuracy
2-out-of-2	0.42261905	0.649133023	0.551403986
1-out-of-2	0.31644669	0.473861284	0.382059466
0-out-of-2	0.03370859	0.168219462	0.066536548

Table 29 An overview of the best and worst accuracies from 4-class classification using soft voting



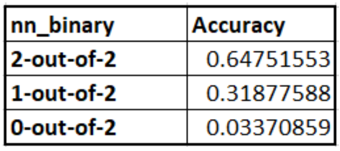
Methods	Min accuracy	Max Accuracy
2-out-of-2	dt_tf	lr_tf
1-out-of-2	lr_tf	rf_binary
0-out-of-2	nn_binary	rf_tf

Table 30 the methods corresponding to the best and worst results from 4-class classification using soft voting



lr_tf	Accuracy
2-out-of-2	0.64913302
1-out-of-2	0.31644669
0-out-of-2	0.03442029

Table 31 LR TF accuracies 4-class classification soft voting

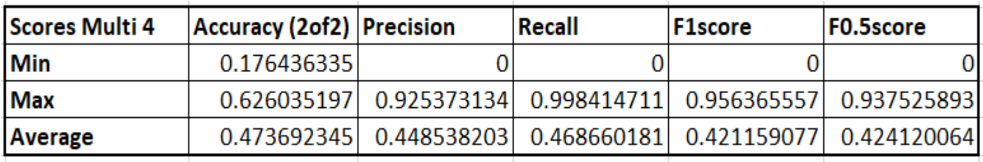


nn_binary	Accuracy
2-out-of-2	0.64751553
1-out-of-2	0.31877588
0-out-of-2	0.03370859

Table 32 NN Binary accuracies 4-class classification soft voting

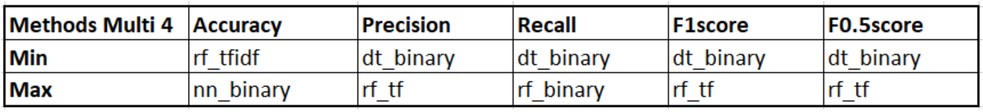
5.2.2. Method II (hard vote)

The results from processing method II can be represented as Table 33 and Table 34 below, along with some of the corresponding confusion matrices: Table 35 and Table 36.



Scores Multi 4	Accuracy (2of2)	Precision	Recall	F1score	F0.5score
Min	0.176436335	0	0	0	0
Max	0.626035197	0.925373134	0.998414711	0.956365557	0.937525893
Average	0.473692345	0.448538203	0.468660181	0.421159077	0.424120064

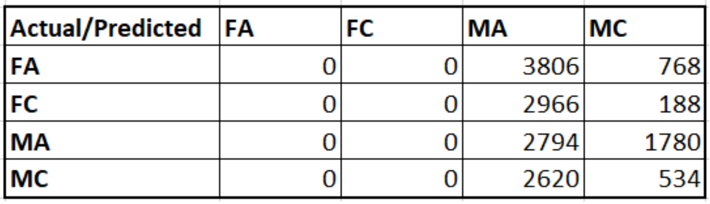
Table 33 The best and worst results from 4-class classification using hard voting



Methods Multi 4	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tfidf	dt_binary	dt_binary	dt_binary	dt_binary
Max	nn_binary	rf_tf	rf_binary	rf_tf	rf_tf

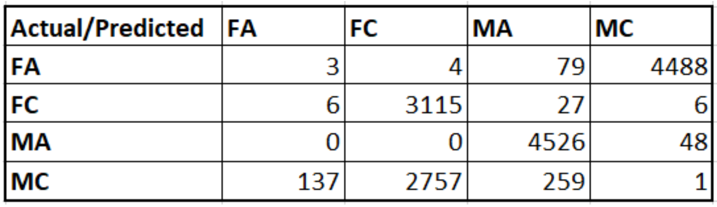
Table 34 the methods corresponding to the best and worst results from 4-class classification using hard voting

The results here are a bit more skewed, as the initial hard voting on the sentences set the highest class to 1, while the rest were set to 0. This is a direct manipulation of the results as the rounding after is very much affected, while initial soft voting is a more indirect approach. DT Binary performs sub optimally here, as shown in Table 35. RF oppositely has very good scores, as shown in Table 36, specifically for classification of FC and MA, along with NN Binary with regards to accuracy.



Actual/Predicted	FA	FC	MA	MC
FA	0	0	3806	768
FC	0	0	2966	188
MA	0	0	2794	1780
MC	0	0	2620	534

Table 35 DT Binary - 4-class classification hard voting

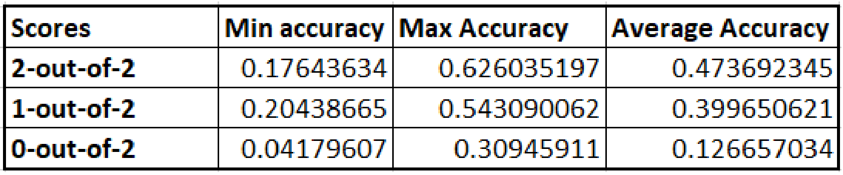


Actual/Predicted	FA	FC	MA	MC
FA	3	4	79	4488
FC	6	3115	27	6
MA	0	0	4526	48
MC	137	2757	259	1

Table 36 RF TF - 4-class classification hard voting

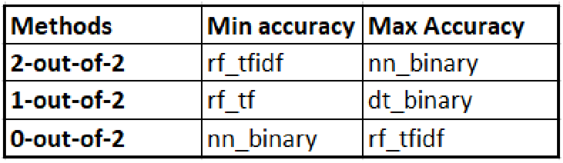
While RF TF and RF Binary performs well with regards to F scores, precision, and recall, it suffers from poor performance when it comes to accuracy, as shown in Table 37 and Table 38 below. It is still not a good classifier as most of the FA is classified as MC, which is undesirable results, and MC is mostly classified as FC.

As Table 39 and Table 40 show, NN Binary performs almost as strong as with soft voting, having a combined accuracy for at least one class correct of almost 96%. RF yields the worst results overall using hard voting.



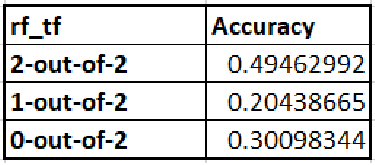
Scores	Min accuracy	Max Accuracy	Average Accuracy
2-out-of-2	0.17643634	0.626035197	0.473692345
1-out-of-2	0.20438665	0.543090062	0.399650621
0-out-of-2	0.04179607	0.30945911	0.126657034

Table 37 An overview of the best and worst accuracies from 4-class classification using hard voting



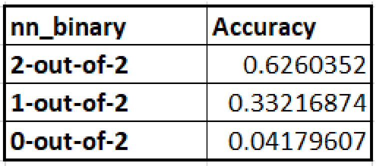
Methods	Min accuracy	Max Accuracy
2-out-of-2	rf_tfidf	nn_binary
1-out-of-2	rf_tf	dt_binary
0-out-of-2	nn_binary	rf_tfidf

Table 38 the methods corresponding to the best and worst results from 4-class classification using hard voting



rf_tf	Accuracy
2-out-of-2	0.49462992
1-out-of-2	0.20438665
0-out-of-2	0.30098344

Table 39 RF TF accuracies 4-class classification hard voting



nn_binary	Accuracy
2-out-of-2	0.6260352
1-out-of-2	0.33216874
0-out-of-2	0.04179607

Table 40 NN Binary accuracies 4-class classification hard voting

5.3. Results of the combined 4-class data classifiers

The 4-class combined data has been processed in the same way as the binary data as shown in Figure 11 below:

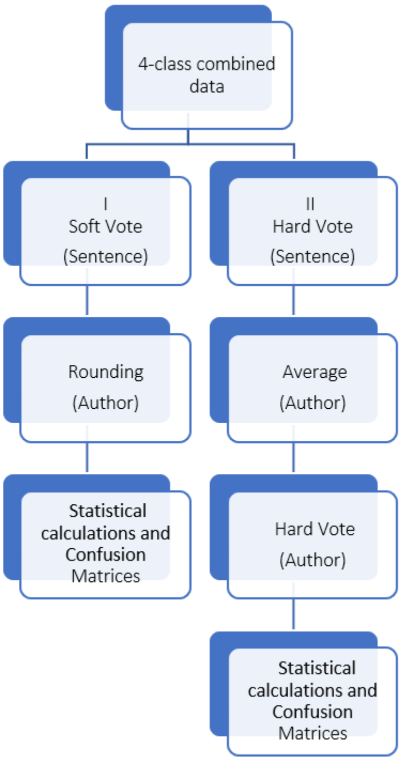


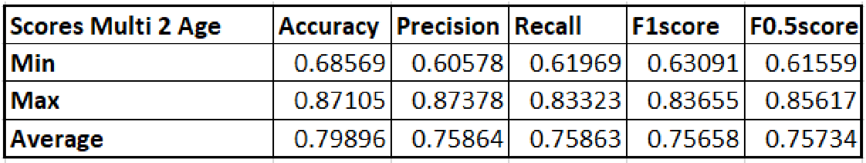
Figure 11 An overview of the processing methods applied to the 4-class combined dataset

The data was combined in two ways in order to best compare these methods to the binary classifications. Firstly, the Male Adult (MA) and Male Children (MC), and Female Adult (FA) and Female Children (FC), were combined into Male and Female classes. Secondly the same was applied to FA and MA, and FC and MC, into Adult and Child classes.

5.3.1. Age classifier

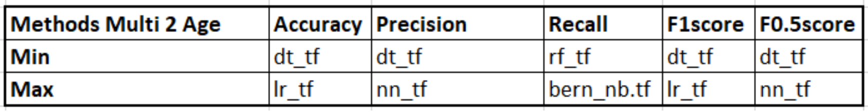
5.3.1.1. Method I (soft vote)

The results from processing method I (soft vote) can be represented as Table 41 and Table 42 below, along with some of the corresponding confusion matrices: Table 43 and Table 44.



Scores Multi 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.68569	0.60578	0.61969	0.63091	0.61559
Max	0.87105	0.87378	0.83323	0.83655	0.85617
Average	0.79896	0.75864	0.75863	0.75658	0.75734

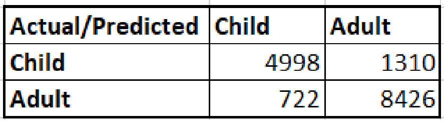
Table 41 The best and worst results from 4-class combined age classification using soft voting



Methods Multi 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	dt_tf	dt_tf	rf_tf	dt_tf	dt_tf
Max	lr_tf	nn_tf	bern_nb.tf	lr_tf	nn_tf

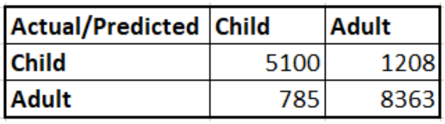
Table 42 the methods corresponding to the best and worst results from 4-class combined age classification using soft voting

There are some more differences with regards to the different methods using this approach. There is no single, or close to a single, method that yields the best scores overall, but NN shows the best results in both precision and $F_{0.5}$ score while LR has the best results accuracy and F_1 score-wise. This is not surprising as NN performed very good on the 4-class classification as well, but this was enabled by using the 4-class classification data as NN was unable to perform well on binary classification data. Therefore, only assumptions can be made about how well it performs compared to the binary classification approach.



Actual/Predicted	Child	Adult
Child	4998	1310
Adult	722	8426

Table 43 NN TF on 4-class combined age data soft vote

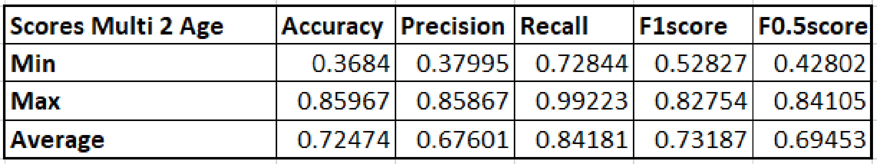


Actual/Predicted	Child	Adult
Child	5100	1208
Adult	785	8363

Table 44 LR TF on 4-class combined age data soft vote

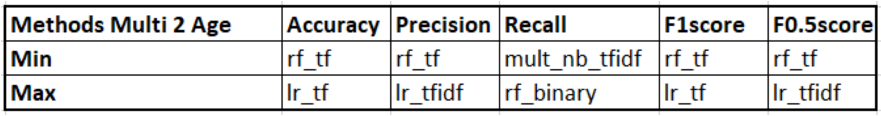
5.3.1.2. Method II (hard vote)

The results from processing method II (hard vote) can be represented as Table 45 and Table 46 below, along with some of the corresponding confusion matrices: Table 47 and Table 48.



Scores Multi 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.3684	0.37995	0.72844	0.52827	0.42802
Max	0.85967	0.85867	0.99223	0.82754	0.84105
Average	0.72474	0.67601	0.84181	0.73187	0.69453

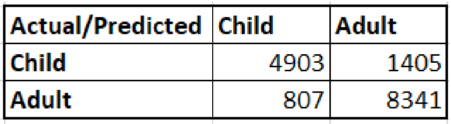
Table 45 The best and worst results from 4-class combined age classification using hard voting



Methods Multi 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tf	rf_tf	mult_nb_tfidf	rf_tf	rf_tf
Max	lr_tf	lr_tfidf	rf_binary	lr_tf	lr_tfidf

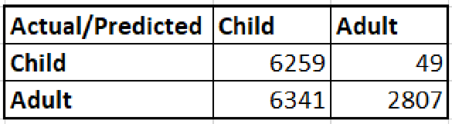
Table 46 the methods corresponding to the best and worst results from 4-class combined age classification using hard voting

RF produces poor results, with the highest precision score attributed to classifying 81% of the authors as children, of which 6341 were adults, as seen in Table 48. LR, which also performed well with the binary classifiers, has the highest scores for most of the metrics. Both the poor performance of RF and good performance of LR a repeat from the binary classification.



Actual/Predicted	Child	Adult
Child	4903	1405
Adult	807	8341

Table 47 LR TF-IDF on 4-class combined age data hard vote



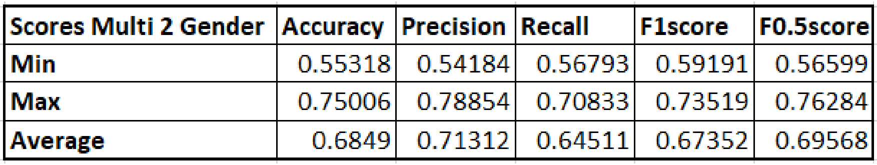
Actual/Predicted	Child	Adult
Child	6259	49
Adult	6341	2807

Table 48 RF Binary on 4-class combined age data hard vote

5.3.2. Gender classifier

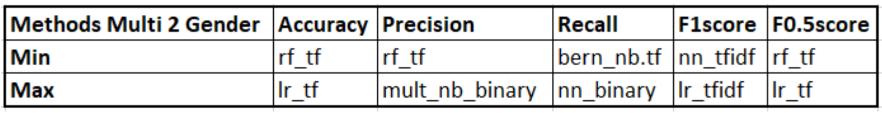
5.3.2.1. Method I (soft vote)

The results from processing method I (soft vote) can be represented as Table 49 and Table 50 below, along with some of the corresponding confusion matrices: Table 51 and Table 52.



Scores Multi 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.55318	0.54184	0.56793	0.59191	0.56599
Max	0.75006	0.78854	0.70833	0.73519	0.76284
Average	0.6849	0.71312	0.64511	0.67352	0.69568

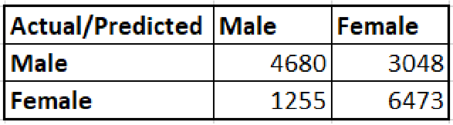
Table 49 The best and worst results from 4-class combined gender classification using soft voting



Methods Multi 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tf	rf_tf	bern_nb.tf	nn_tfidf	rf_tf
Max	lr_tf	mult_nb_binary	nn_binary	lr_tfidf	lr_tf

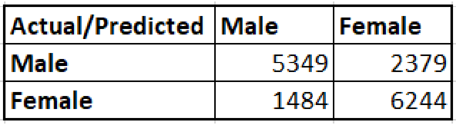
Table 50 the methods corresponding to the best and worst results from 4-class combined gender classification using soft voting

RF keeps producing poor results. LR still shows how it is applicable to this type of data and classification, while NN has the best recall instead of precision as it had with age classification. The difference in gender and age classification results, as with the binary classification, becomes apparent as both the minimum scores are lower, and the maximum scores and average scores are higher for the age classification.



Actual/Predicted	Male	Female
Male	4680	3048
Female	1255	6473

Table 51 Multinomial NB Binary on 4-class combined gender data soft vote

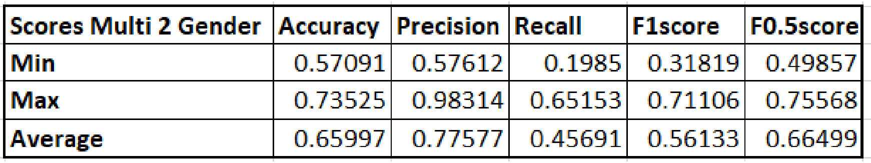


Actual/Predicted	Male	Female
Male	5349	2379
Female	1484	6244

Table 52 LR TF on 4-class combined gender data soft vote

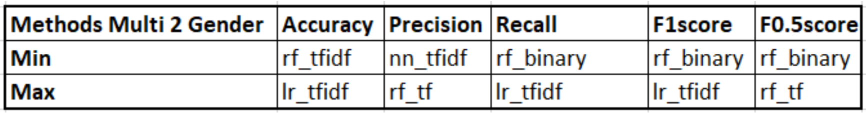
5.3.2.2. Method II (hard vote)

The results from processing method II (hard vote) can be represented as Table 53 and Table 54 below, along with some of the corresponding confusion matrices: Table 55 and Table 56.



Scores Multi 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.57091	0.57612	0.1985	0.31819	0.49857
Max	0.73525	0.98314	0.65153	0.71106	0.75568
Average	0.65997	0.77577	0.45691	0.56133	0.66499

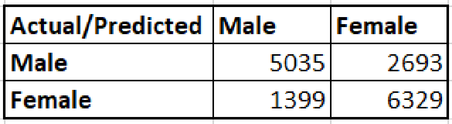
Table 53 The best and worst results from 4-class combined gender classification using hard voting



Methods Multi 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tfidf	nn_tfidf	rf_binary	rf_binary	rf_binary
Max	lr_tfidf	rf_tf	lr_tfidf	lr_tfidf	rf_tf

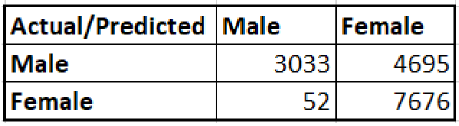
Table 54 the methods corresponding to the best and worst results from 4-class combined gender classification using hard voting

As with all the results where RF is on top, the classification has in reality very poor results, here having the best precision while classifying most authors as female, as seen in Table 56. LR still shows good results and RF still has poor results overall.



Actual/Predicted	Male	Female
Male	5035	2693
Female	1399	6329

Table 55 LR TF-IDF on 4-class combined age data hard vote



Actual/Predicted	Male	Female
Male	3033	4695
Female	52	7676

Table 56 RF TF on 4-class combined age data hard vote

As with the precision scores using soft voting, the age classification yielded better results in 13 out of 18 methods. The RF methods especially, and Bern NB and DT have the differences amplified by the hard voting, while the other methods seem to be more adjusted towards similar results between age and gender classification.

The distribution of $F_{0.5}$ scores for both the age and gender classifiers for the combined data are as shown below in Figure 12 and Figure 13. Again, the results for $F_{0.5}$ scores are the inverse from precision hard voting. The $F_{0.5}$ scores show that age classification is highest in 11 out of 18 methods, while precision only comes out on top on 7 out of 18 methods using hard voting.

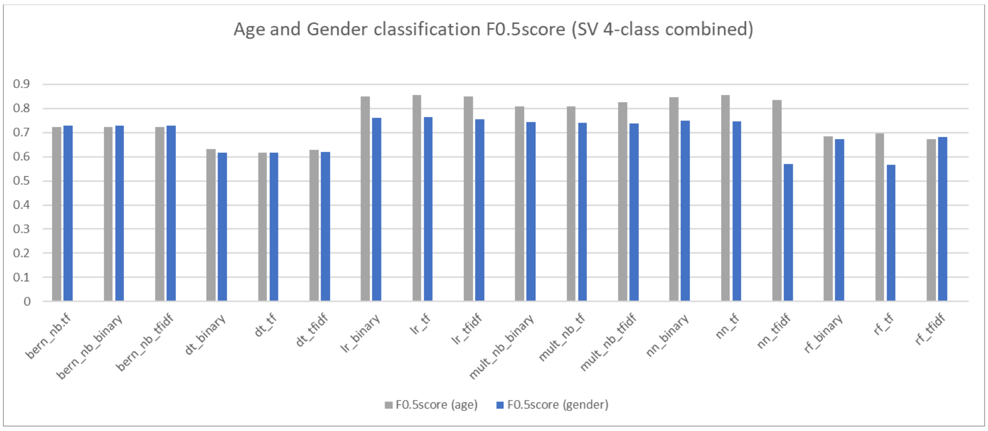


Figure 12 $F_{0.5}$ scores for Age and Gender classification 4-class combined input data soft vote

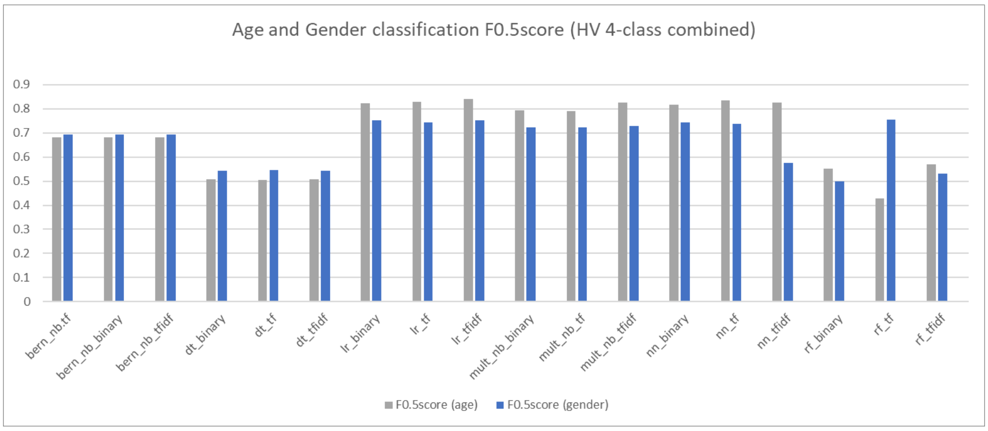


Figure 13 F0.5 scores for Age and Gender classification 4-class combined input data hard vote

5.4. Comparison and discussion

The scores from soft and hard voting seem to have repeating patterns. The sentence-based hard voting is a bit unforgiving, not only for the 4-class classification, which yielded subpar results overall without any adjustments compared to the other approaches, but also for the binary and combined datasets. The results are more equalized between age and gender, but also lower on an overall view compared to sentence-based soft voting. It has also become evident that age classification has produced better results than gender classification.

Figure 14-Figure 17 below show the different methods using binary input, both the binary data and the combined 4-class data. The 4-class combined methods outperform the binary methods for most of the algorithms. LR, which is the best performing method overall, has quite similar scores for both 4-class combined classification and binary classification. The difference is so small it is negligible. The methods with lower performance have more considerable differences, varying between which is better, 4-class combined or binary classification.

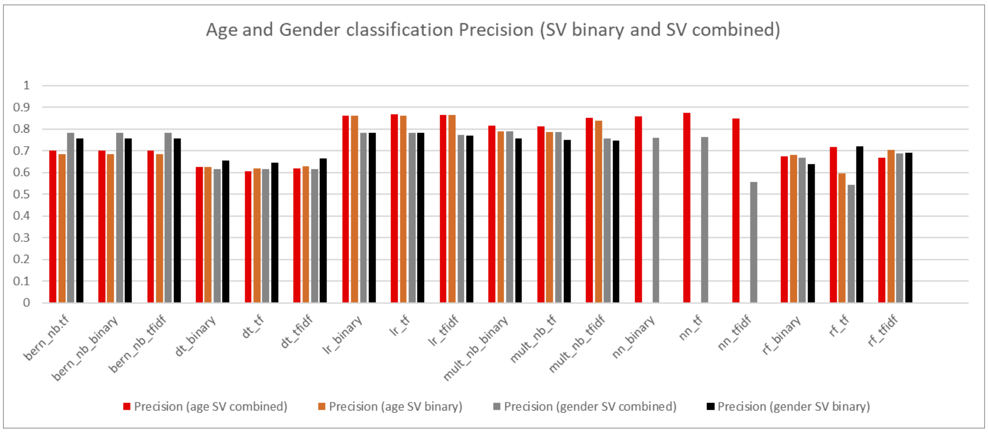


Figure 14 Precision scores for Age and Gender classification Binary and 4-class combined input data soft vote

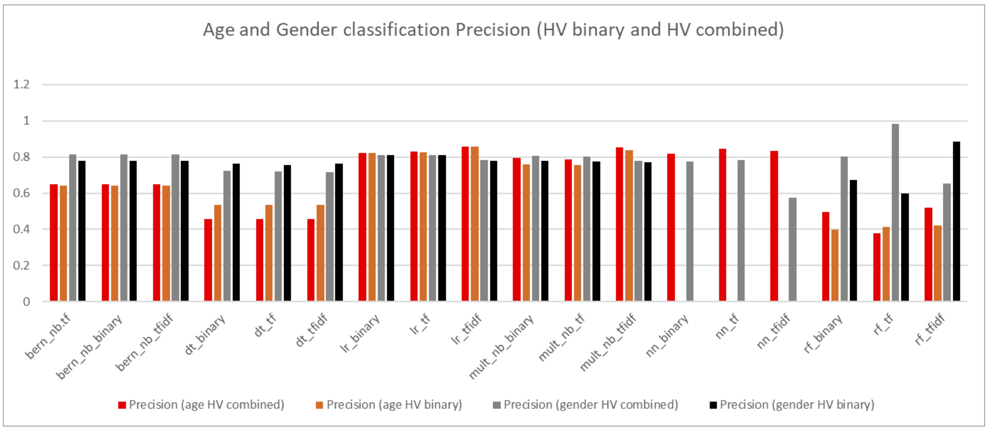


Figure 15 Precision scores for Age and Gender classification Binary and 4-class combined input data hard vote

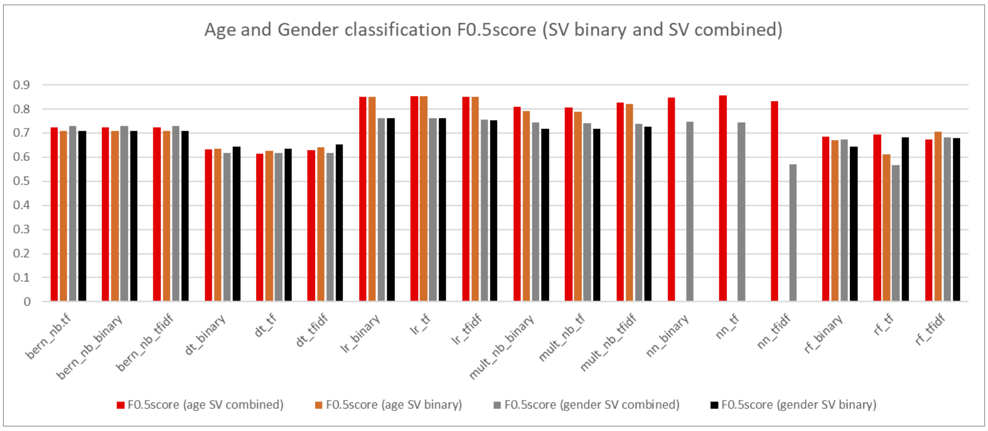


Figure 16 F0.5 scores for Age and Gender classification Binary and 4-class combined input data soft vote

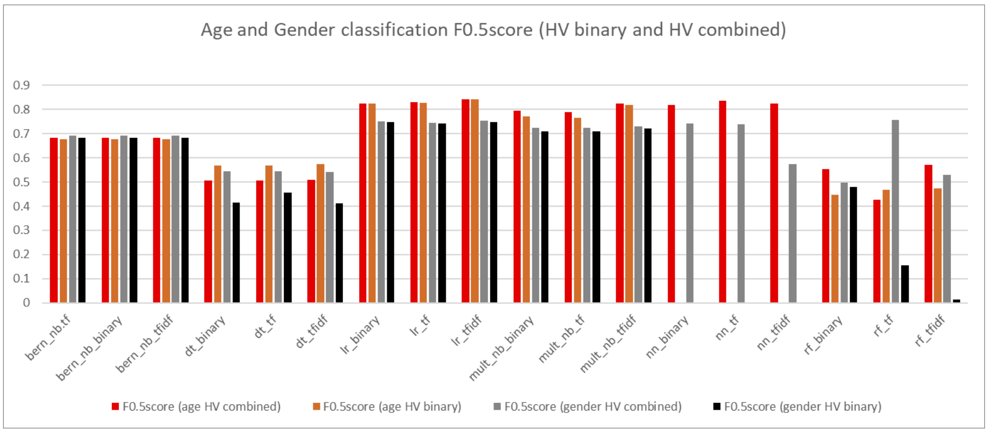
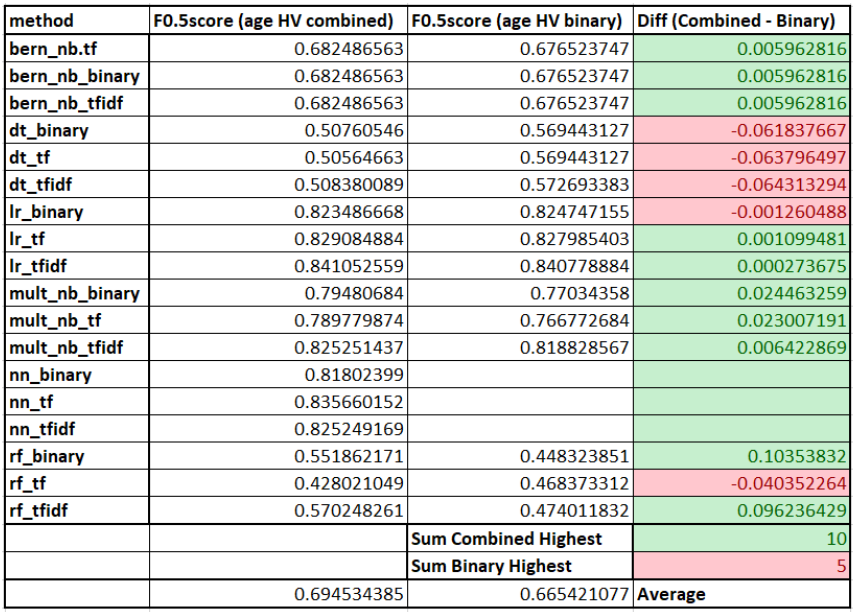


Figure 17 $F_{0.5}$ scores for Age and Gender classification Binary and 4-class combined input data hard vote

As Table 57-Table 60 below show, the methods where binary classification show better results are also often the methods with the lowest scores. NN has been removed from the comparison of how many methods of which 4-class classification is better because the algorithm showed unusable data on the binary classification results. NN often showed notable results using 4-class classification, combined or not. The average score at the bottom includes NN for the 4-class combined results to provide a more correct score. These numbers present the overall score of how good each method is, but it must be emphasized that some of the differences are very small. Precision scores can be found in the appendix.

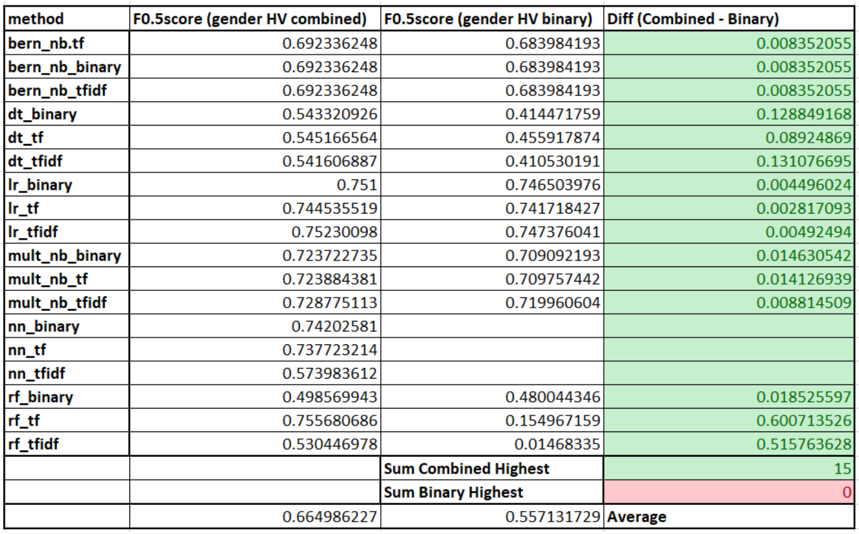
The total of occurrences, for all methods including algorithms, feature extraction methods, soft and hard voting, and 4-class combined and binary classification, excluding NN, where the 4-class combined have better results within precision and $F_{0.5}$ score is 85 out of a total of 120, meaning 70.8%. This implies that 4-class combined classification is better than two binary classifications, which is not in line with (Ferdous, et al., 2020)'s assumption. It does however substantiate the results of (Peersman, et al., 2011), who also achieved better results using what they called a 4-way classification rather than binary classifications. Even if the methods they used differ, and the dataset is different, they also emphasize another finding of this thesis, that differences in language is more substantial with regards to age rather than gender.

Table 58 stands out, where 4-class combined classification outperforms binary classification for every method in $F_{0.5}$ score with gender classification hard vote. The difference between the methods ranges from negligible to 0.6, with the average difference in precision and $F_{0.5}$ scores is as showed in Table 61 below. The difference there varies from negligible, to 10.7%, and is on an average in favor of the combined 4-class classification with 1.6% and 4%, precision and $F_{0.5}$ scores respectively.



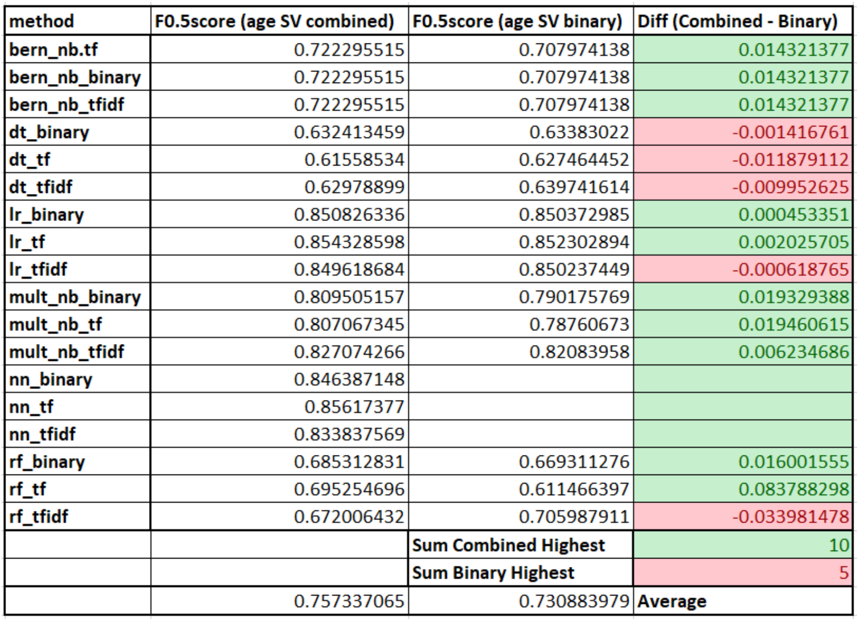
method	F0.5score (age HV combined)	F0.5score (age HV binary)	Diff (Combined - Binary)
bern_nb.tf	0.682486563	0.676523747	0.005962816
bern_nb_binary	0.682486563	0.676523747	0.005962816
bern_nb_tfidf	0.682486563	0.676523747	0.005962816
dt_binary	0.50760546	0.569443127	-0.061837667
dt_tf	0.50564663	0.569443127	-0.063796497
dt_tfidf	0.508380089	0.572693383	-0.064313294
lr_binary	0.823486668	0.824747155	-0.001260488
lr_tf	0.829084884	0.827985403	0.001099481
lr_tfidf	0.841052559	0.840778884	0.000273675
mult_nb_binary	0.79480684	0.77034358	0.024463259
mult_nb_tf	0.789779874	0.766772684	0.023007191
mult_nb_tfidf	0.825251437	0.818828567	0.006422869
nn_binary	0.81802399		
nn_tf	0.835660152		
nn_tfidf	0.825249169		
rf_binary	0.551862171	0.448323851	0.10353832
rf_tf	0.428021049	0.468373312	-0.040352264
rf_tfidf	0.570248261	0.474011832	0.096236429
		Sum Combined Highest	10
		Sum Binary Highest	5
	0.694534385	0.665421077	Average

Table 57 A comparison of the F0.5 scores of all methods within 4-class combined and binary age classification hard vote



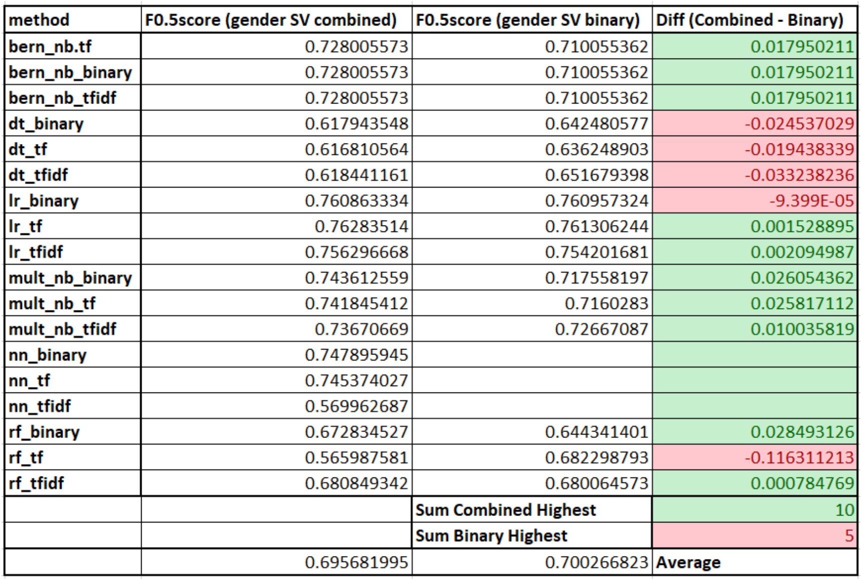
method	F0.5score (gender HV combined)	F0.5score (gender HV binary)	Diff (Combined - Binary)
bern_nb.tf	0.692336248	0.683984193	0.008352055
bern_nb_binary	0.692336248	0.683984193	0.008352055
bern_nb_tfidf	0.692336248	0.683984193	0.008352055
dt_binary	0.543320926	0.414471759	0.128849168
dt_tf	0.545166564	0.455917874	0.08924869
dt_tfidf	0.541606887	0.410530191	0.131076695
lr_binary	0.751	0.746503976	0.004496024
lr_tf	0.744535519	0.741718427	0.002817093
lr_tfidf	0.75230098	0.747376041	0.00492494
mult_nb_binary	0.723722735	0.709092193	0.014630542
mult_nb_tf	0.723884381	0.709757442	0.014126939
mult_nb_tfidf	0.728775113	0.719960604	0.008814509
nn_binary	0.74202581		
nn_tf	0.737723214		
nn_tfidf	0.573983612		
rf_binary	0.498569943	0.480044346	0.018525597
rf_tf	0.755680686	0.154967159	0.600713526
rf_tfidf	0.530446978	0.01468335	0.515763628
		Sum Combined Highest	15
		Sum Binary Highest	0
	0.664986227	0.557131729	Average

Table 58 A comparison of the F0.5 scores of all methods within 4-class combined and binary gender classification hard vote



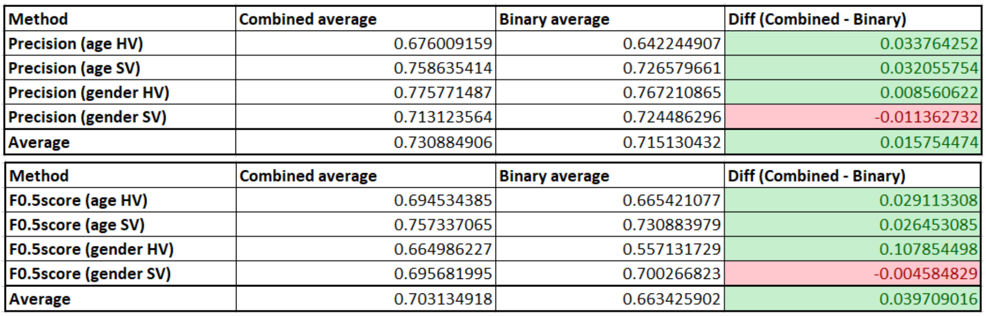
method	F0.5score (age SV combined)	F0.5score (age SV binary)	Diff (Combined - Binary)
bern_nb.tf	0.722295515	0.707974138	0.014321377
bern_nb_binary	0.722295515	0.707974138	0.014321377
bern_nb_tfidf	0.722295515	0.707974138	0.014321377
dt_binary	0.632413459	0.63383022	-0.001416761
dt_tf	0.61558534	0.627464452	-0.011879112
dt_tfidf	0.62978899	0.639741614	-0.009952625
lr_binary	0.850826336	0.850372985	0.000453351
lr_tf	0.854328598	0.852302894	0.002025705
lr_tfidf	0.849618684	0.850237449	-0.000618765
mult_nb_binary	0.809505157	0.790175769	0.019329388
mult_nb_tf	0.807067345	0.78760673	0.019460615
mult_nb_tfidf	0.827074266	0.82083958	0.006234686
nn_binary	0.846387148		
nn_tf	0.85617377		
nn_tfidf	0.833837569		
rf_binary	0.685312831	0.669311276	0.016001555
rf_tf	0.695254696	0.611466397	0.083788298
rf_tfidf	0.672006432	0.705987911	-0.033981478
		Sum Combined Highest	10
		Sum Binary Highest	5
	0.757337065	0.730883979	Average

Table 59 A comparison of the F0.5 scores of all methods within 4-class combined and binary age classification soft vote



method	F0.5score (gender SV combined)	F0.5score (gender SV binary)	Diff (Combined - Binary)
bern_nb.tf	0.728005573	0.710055362	0.017950211
bern_nb_binary	0.728005573	0.710055362	0.017950211
bern_nb_tfidf	0.728005573	0.710055362	0.017950211
dt_binary	0.617943548	0.642480577	-0.024537029
dt_tf	0.616810564	0.636248903	-0.019438339
dt_tfidf	0.618441161	0.651679398	-0.033238236
lr_binary	0.760863334	0.760957324	-9.399E-05
lr_tf	0.76283514	0.761306244	0.001528895
lr_tfidf	0.756296668	0.754201681	0.002094987
mult_nb_binary	0.743612559	0.717558197	0.026054362
mult_nb_tf	0.741845412	0.7160283	0.025817112
mult_nb_tfidf	0.73670669	0.72667087	0.010035819
nn_binary	0.747895945		
nn_tf	0.745374027		
nn_tfidf	0.569962687		
rf_binary	0.672834527	0.644341401	0.028493126
rf_tf	0.565987581	0.682298793	-0.116311213
rf_tfidf	0.680849342	0.680064573	0.000784769
		Sum Combined Highest	10
		Sum Binary Highest	5
	0.695681995	0.700266823	Average

Table 60 A comparison of the F0.5 scores of all methods within 4-class combined and binary gender classification soft vote



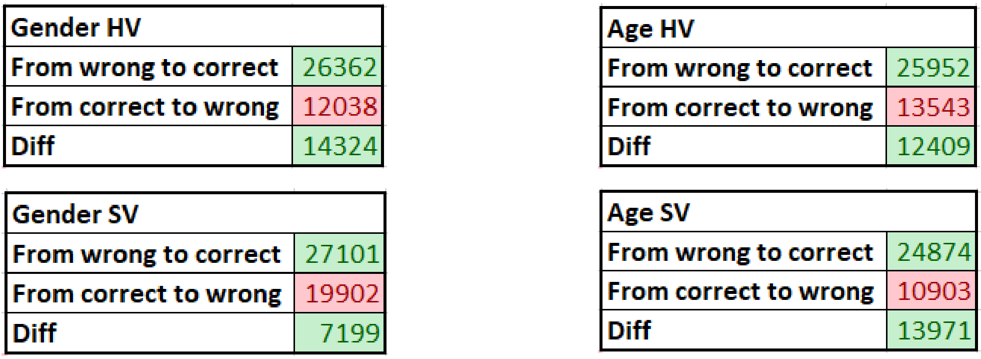
Method	Combined average	Binary average	Diff (Combined - Binary)
Precision (age HV)	0.676009159	0.642244907	0.033764252
Precision (age SV)	0.758635414	0.726579661	0.032055754
Precision (gender HV)	0.775771487	0.767210865	0.008560622
Precision (gender SV)	0.713123564	0.724486296	-0.011362732
Average	0.730884906	0.715130432	0.015754474

Method	Combined average	Binary average	Diff (Combined - Binary)
F0.5score (age HV)	0.694534385	0.665421077	0.029113308
F0.5score (age SV)	0.757337065	0.730883979	0.026453085
F0.5score (gender HV)	0.664986227	0.557131729	0.107854498
F0.5score (gender SV)	0.695681995	0.700266823	-0.004584829
Average	0.703134918	0.663425902	0.039709016

Table 61 Average precision and $F_{0.5}$ scores compared

Another interesting metric is how many of the different authors were classified differently between 4-class combined classification and binary classification. Table 62 below show many were changed to correct classification, and how many were changed to incorrect classification, from binary classification to 4-class combined classification.

It is important to note that even though there are only 15456 authors, these differences are from all the final classifications of the authors, meaning that for each table above, there have been 18 times more classifications than authors. Therefore, the difference between the number of classifications that has been corrected and those that have changed to a wrong classification, is out of 278208 classifications. Even so, for most of the methods, there have been over 10000 more correct classifications for the 4-class combined classification than for the binary classification. This means that the 4-class combined classification classified on average 4.3% more correctly, as Table 63 below shows



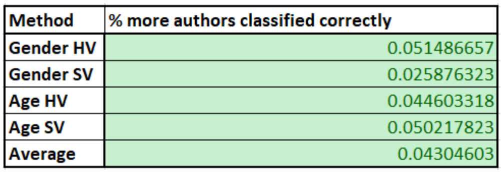
Gender HV	
From wrong to correct	26362
From correct to wrong	12038
Diff	14324

Age HV	
From wrong to correct	25952
From correct to wrong	13543
Diff	12409

Gender SV	
From wrong to correct	27101
From correct to wrong	19902
Diff	7199

Age SV	
From wrong to correct	24874
From correct to wrong	10903
Diff	13971

Table 62 Differences in classification between 4-class and binary classification

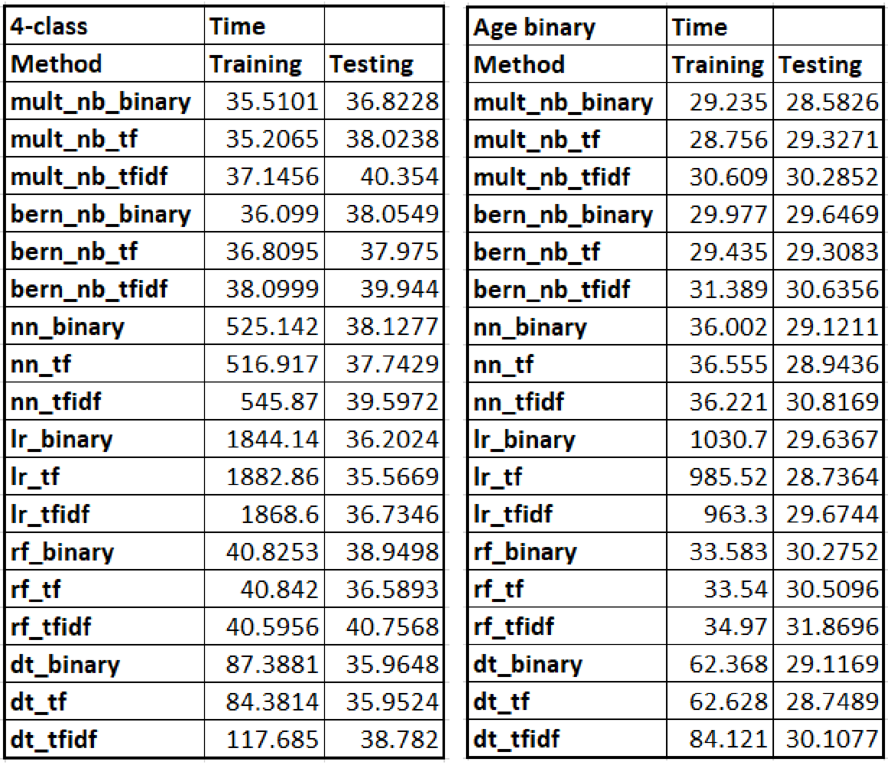


Method	% more authors classified correctly
Gender HV	0.051486657
Gender SV	0.025876323
Age HV	0.044603318
Age SV	0.050217823
Average	0.04304603

Table 63 Difference in correctly classified authors

There are differences between 4-class classification and binary classification other than the performance of the classification. Table 64 show the computing times for all the algorithms, both for the training and the testing part for 4-class classification and binary classification, performed by Fauzi. The computing times for binary gender classification can be found in section A.4 of the appendix and is very similar to the computing times for the binary age classification.

While the 4-class classification clearly is somewhat slower, especially the training part, it must only run once, instead of twice, both for age and gender classification, which the binary classification needs. LR specifically needs longer training times than the rest of the algorithms, even more so for the 4-class classification, but the training must only be done once, or less frequent than the testing, for both binary and 4-class classification. By this logic the 4-class classification outperforms the binary classification when it comes to computing times, cutting the time required for testing almost in half.



4-class	Time	
Method	Training	Testing
mult_nb_binary	35.5101	36.8228
mult_nb_tf	35.2065	38.0238
mult_nb_tfidf	37.1456	40.354
bern_nb_binary	36.099	38.0549
bern_nb_tf	36.8095	37.975
bern_nb_tfidf	38.0999	39.944
nn_binary	525.142	38.1277
nn_tf	516.917	37.7429
nn_tfidf	545.87	39.5972
lr_binary	1844.14	36.2024
lr_tf	1882.86	35.5669
lr_tfidf	1868.6	36.7346
rf_binary	40.8253	38.9498
rf_tf	40.842	36.5893
rf_tfidf	40.5956	40.7568
dt_binary	87.3881	35.9648
dt_tf	84.3814	35.9524
dt_tfidf	117.685	38.782

Age binary	Time	
Method	Training	Testing
mult_nb_binary	29.235	28.5826
mult_nb_tf	28.756	29.3271
mult_nb_tfidf	30.609	30.2852
bern_nb_binary	29.977	29.6469
bern_nb_tf	29.435	29.3083
bern_nb_tfidf	31.389	30.6356
nn_binary	36.002	29.1211
nn_tf	36.555	28.9436
nn_tfidf	36.221	30.8169
lr_binary	1030.7	29.6367
lr_tf	985.52	28.7364
lr_tfidf	963.3	29.6744
rf_binary	33.583	30.2752
rf_tf	33.54	30.5096
rf_tfidf	34.97	31.8696
dt_binary	62.368	29.1169
dt_tf	62.628	28.7489
dt_tfidf	84.121	30.1077

Table 64 Computing times for 4-class and binary age classification in seconds

Comparing the results from the 4-class combined data classification with the most relevant related works, as presented in Table 1, it appears that this study does not measure up to their highest scores. (Holbæk, 2019) achieved an F_1 score of 0.90 with linear SVM. The only method used in both studies was Bernoulli NB, which yielded almost identical results. This study achieved 0.76 with Bernoulli NB, and (Holbæk, 2019) 0.77. There are differences in the methods, as e.g., LIWC, n-grams and PoS which were implemented in that work. (Holbæk, 2019) also used the entire text for each author for a classification, while the approach of this thesis is more prepared for the continuous classification approach, classifying every sentence per author. (Peersman, et al., 2011)'s research did not use the same dataset as this but achieved an F_1 score of 0.917. This study, when using the combined data, achieved the highest score using soft voting, with an F_1 score of 0.836 and an $F_{0.5}$ score of 0.856 on age classification, with LR TF and NN TF respectively. The difference in score values could be attributed to the difference in methods and dataset, where (Peersman, et al., 2011) used chat data, trained the classifiers with the most informative unigram features, and adjusted the size of the dataset.

Comparing the achieved scores from the classifications with other related work is both difficult and not the focus of this study. The accuracy of the results is therefore only important for comparing the binary and 4-class combined approach, not for comparing with another research. Using the same approach of four classes instead of two binary classes, with identical methods as the related work would be a better comparison. The four classes are also very much linked to each other, which made it hard to find other studies with similar classification problems.

6. Conclusion and future work

6.1. Conclusion

Firstly, when the classifier models were trained and classified, the computing times were recorded, see Table 64. It became apparent that the computing times for the 4-class classifiers were slightly higher, especially for the more CPU-heavy methods such as LR. Even so, considering that the 4-class classifier must only run once compared to twice for the two binary classifiers, age and gender, the 4-class classifiers is by far the faster choice.

Secondly, the results from the classifications per sentence were processed through both soft and hard voting. The soft voting yielded higher precision and $F_{0.5}$ scores overall, as the hard voting was a bit unforgiving, as explained in section 5.4. For the results to be comparable, the results from the 4-class classification had to be combined age and gender-wise, as to match the format of the binary classification results. The results from the 4-class classification without combining the scores performed subpar compared to the binary classification, especially for hard voting in how it was implemented.

Thirdly, by using voting and combining the 4-class classification results, the results, as presented in Table 57-Table 60, showed that the combination of the 4-class classification results performed better in 70.8% of all the different methods with regards to precision and $F_{0.5}$ scores. The difference per method varies from negligible to up to 60% for the worst performing methods overall. The average difference in performance for precision and $F_{0.5}$ score was not more than 1.6-4%, but it did classify on average 4.3% more authors correctly, as shown in Table 61 and Table 63.

This study was mainly conducted to find out whether classification of both age and gender through text analysis and machine learning differ when treating it as a 4-class problem rather than two binary problems or not. The experiment was organized so that every algorithm, feature set and type of voting could be compared with each other. Not necessarily looking at the combination of all methods, as would be an ensemble method. Based on these comparisons, it is a notable difference in favor of the 4-class classification, which is not in line with the assumption made by (Ferdous, et al., 2020) that binary methods are more accurate in age and gender classification, but substantiates the findings of (Peersman, et al., 2011). It is difficult to compare these results to the results of another research with different methods, e.g., different algorithms, looking at full texts instead of sentences, or other alternatives.

6.2. Future work

Regarding future work there are several areas that could be explored. Ensemble method is known to improve the performance of binary classifiers, which could be applicable to 4-class classification as well. It should also be explored how this method applies to other classification problems than specifically age and gender, such as gender and native/non-native English speaker.

Another area could be introducing another type of dataset. If the dataset is more comprised of actual chat data, it should be more relevant for the “chatroom security” research at NTNU. The experiment conducted in this study could also be applied to other languages, in order to incorporate a more global approach. The next languages could therefore be split into two groups, one for widely spoken western languages with similar structure to English, such as Spanish, French, and German, and other major languages structured differently, such as Mandarin and Arabic.

This thesis did not use all the known, or most popular algorithms used for age and gender classification based on text. Some specific algorithms that should also be tested with the same experiment is k-nearest neighbors and SVM, which have already proven themselves popular in this field of research. This could be in combination with using other methods than, or an addition to, voting, such as a sigmoid function or only using values within certain thresholds. This could potentially increase the performance for the 4-class classification data, as the hard voting had quite an impact on the results. This study only used BoW feature sets, which could limit the results. It could be interesting to look at word embeddings, keystroke dynamics, average word and sentence length, number of questions, PoS and other popular methods to see what kind of results are produced using the same approach. This could be especially applicable for comparison if this approach is applied to specific methods of another research.

The last suggestion for future work in this study is looking at e.g., how often, in detection of age and gender in real time context such as chatrooms, should the algorithms be re-trained to incorporate new data. Some algorithms such as artificial neural networks do not need to be re-trained, or at least not to include new data, which could be the more relevant approach, especially for the 4-class classifier which yielded very good results for NN. A supplement to this suggestion could be to look at the difference and performance of different NN, like conventional, artificial, and recurrent NN. Also, how early can a decision be made as to what class the author belongs to, how much data is needed, and how could the algorithms used in this study and the methods used be adjusted to these challenges.

Bibliography

Amble, Z., 2021. *TV2.* [Online]
Available at: https://play.tv2.no/programmer/fakta/norge-bak-fasaden/sesong-2/norge-bak-fasaden-2-episode-3-1622683.html

Banerjee, R., Feng, S., Kang, J. S. & Choi, Y., 2014. *Keystroke Patterns as Prosody in Digital Writings: A Case Study with Deceptive Reviews and Essays.* Doha, Association for Computational Linguistics, pp. 1469-1473.

Bello, H. R. M., Heilmann, L. & Ronan, E., 2020. *Detecting Fake News Spreaders with Behavioural, Lexical and Psycholinguistic Features.* Thessaloniki, CEUR.

Bevendorff, J. et al., 2020. *Shared Tasks on Authorship Analysis at PAN 2020.* s.l., Springer, Cham, pp. 508-516.

Bond Jr, C. F. & DePaulo, B. M., 2006. Accuracy of Deception Judgments. *Personality and Social Psychology Review,* 10(3), pp. 214-234.

Borj, P. R. & Bours, P., 2019. *Predatory Conversation Detection.* Doha, IEEE, pp. 1-6.

Brownlee, J., 2017. *A Gentle Introduction to the Bag-of-Words Model.* [Online]
Available at: https://machinelearningmastery.com/gentle-introduction-bag-words-model/
[Accessed 29 11 2021].

Bunge, J. A. & Judson, D. H., 2005. Data Mining. In: *Encyclopedia of Social Measurement.* s.l.:Elsevier, pp. 617-624.

Burnett, C., 2006. *Wikipedia.* [Online]
Available at:
https://no.m.wikipedia.org/wiki/Fil:Artificial_neural_network.svg
[Accessed 29 11 2021].

Burrows, J., 2007. All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing,* 22(1), pp. 27-47.

Chakure, A., 2019. *Random Forest Regression.* [Online]
Available at: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f
[Accessed 29 11 2021].

Dramsch, J. S., 2020. 70 years of machine learning in geoscience in review. In: *Advances in Geophysics.* s.l.:Elsevier, pp. 1-55.

Edgar, T. W. & Manz, D. O., 2017. *Research Methods for Cyber Security.* 1st ed. s.l.:Elsevier.

Epp, C., Lippold, M. & Mandryk, R. L., 2011. *Identifying Emotional States Using Keystroke Dynamics.* Vancouver, Association for Computing Machinery, pp. 715-724.

Fauzi, M. A., 2018. Automatic complaint classification system using classifier ensembles. *Telfor,* 10(2), pp. 123-128.

Fauzi, M. A. & Bours, P., 2020. *Ensemble Method for Sexual Predators Identification in Online Chats.* Porto, IEEE.

Ferdous, J., Turzo, N. A. & Sarker, P., 2020. Learning Practice for Binary Age Group Stratification by Iris-Pupil Thickness. *International Journal of Scientific & Engineering Research,* 11(7), pp. 92-96.

Gedeck, P., Kramer, C. & Ertl, P., 2010. Computational Analysis of Structure–Activity Relationships. In: *Progress in Medicinal Chemistry.* s.l.:Elsevier, pp. 113-160.

Gudivada, V. N., Raghavan, V. V., Govindaraju, V. & Rao, C. R., 2016. Handbook of Statistics. In: *Cognitive Computing: Theory and Applications.* s.l.:Elsevier, pp. 2-384.

Hancock, J. T., Thom-Santelli, J. & Ritchie, T., 2004. *Deception and design: The impact of communication technology on lying behavior.* Vienna, Conference on Computer Human Interaction, pp. 129-134.

Holbæk, E., 2019. *Using Author Profiling to Determine the Age Group of an Author,* Gjøvik: NTNU.

Holle, K. F. H., Arifin, A. Z. & Purwitasari, D., 2015. Preference based term weighting for arabic fiqh document ranking. *Journal of Computer Science and Information,* 8(1), pp. 45-52.

Huisman, A. B., 2016. *Deception detection using keystroke dynamics,* Zwolle: Universiteit Twente.

Inches, G. & Crestani, F., 2012. *Overview of the International Sexual Predator Identification Competition at PAN-2012.* Rome, s.n.

Kalch, A. & Naab, T., 2017. Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *Studies in Communication and Media,* 6(4), pp. 395-419.

Kittler, J., 2002. Multiple Classifier Systems. In: A. Ghosh & S. K. Pal, eds. *Multiple Classifier Systems.* s.l.:World Scientific, pp. 3-22.

Kotu, V. & Deshpande, B., 2015. Classification. In: *Predictive Analytics and Data Mining.* s.l.:Elsevier, pp. 63-163.

Kowsari, K. et al., 2020. *Gender Detection on Social Networks using Ensemble Deep Learning.* Vancouver, Springer, pp. 346-358.

Kulsrud, H., 2019. *Detection of cyber grooming during an online conversation,* Gjøvik: NTNU.

Larkley, L. & Croft, B. W., 1996. *Combining Classifiers in Text Categorization.* s.l., SIGIR, pp. 289-297.

Luyckx, K. & Daelemans, W., 2010. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing,* 26(1), pp. 35-55.

Marini, F., 2009. Neural Networks. In: S. D. Brown, R. Tauler & B. Walczak, eds. *Comprehensive Chemometrics.* s.l.:Elsevier, pp. 477-505.

McCornack, S. A., 1997. The Generation of Deceptive Messages: Laying the Groundwork for a Viable Theory of Interpersonal Deception. In: J. O. Greene, ed. *Message Production Advances in Communication Theory.* s.l.:Lawrence Erlbaum Associates Publishers, p. 91–126.

Miner, G. et al. eds., 2012. Prediction in Text Mining: The Data Mining Algorithms of Predictive Analytics. In: *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* s.l.:Elsevier, pp. 893-919.

Misra, S. & Li, H., 2020. Noninvasive fracture characterization based on the classification of sonic wave travel times. In: *Machine Learning for Subsurface Characterization.* s.l.:Elsevier, pp. 243-287.

Mushtaq, M.-S. & Mellouk, A., 2017. Methodologies for Subjective Video Streaming QoE Assessment. In: *Quality of Experience Paradigm in Multimedia Services.* s.l.:Elsevier, pp. 27-57.

Navilani, A., 2019. *Understanding Logistic Regression in Python.* [Online]
Available at:
https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python
[Accessed 29 11 2021].

Newman, M. L., Pennebaker, J. W., Berry, D. S. & Richards, J. M., 2003. Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin,* 29(5), pp. 665-675.

Ott, M., Choi, Y., Cardie, C. & Hancock, J. T., 2011. *Finding Deceptive Opinion Spam by Any Stretch of the Imagination.* Portland, Association for Computational Linguistics, pp. 309-319.

Oxford University Press, 2021. *Definition of overfitting.* [Online]
Available at: https://www.lexico.com/definition/overfitting
[Accessed 29 11 2021].

Papacharissi, Z., 2004. Democracy Online: Civility, Politeness, and the Democratic Potential of Online Political Discussion Groups. *New Media & Society,* 6(2), pp. 259-283.

Peersman, C., Daelemans, W. & Vaerenbergh, L. V., 2011. *Predicting age and gender in online social networks.* Glasgow, Association for Computing Machinery, pp. 37-44.

Pendar, N., 2007. *Toward Spotting the Pedophile Telling victim from predator in text chats.* Irvine, IEEE Computer Society Press, pp. 235-241.

Reinders, C., Ackermann, H., Yang, M. Y. & Rosenhahn, B., 2019. Learning Convolutional Neural Networks for Object Detection with Very Little Training Data. In: M. Y. Yang, B. Rosenhahn & V. Murino, eds. *Multimodal Scene Understanding.* s.l.:Elsevier, pp. 65-100.

Sanderson, C. & Guenter, S., 2006. *Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation.* Sydney, Association for Computational Linguistics, pp. 482-491.

Schler, J., Koppel, M., Argamon, S. & Pennebaker, J. W., 2006. *Effects of Age and Gender on Blogging.* Stanford, AAAI.

Seufert, E. B., 2014. Quantitative Methods for Product Management. In: *Freemium Economics.* 1st ed. s.l.:Elsevier, pp. 47-82.

Silva, J. et al., 2020. *A method for detecting the profile of an author.* Warsaw, Procedia Computer Science, pp. 959-964.

Singh, G., Bhawna, K., Gaur, L. & Tyagi, A., 2019. *Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification.* s.l., IEEE, pp. 593-596.

Singh, V., 2018. *Machine Learning Logistic Regression In Python: From Theory To Trading.* [Online]
Available at: https://blog.quantinsti.com/machine-learning-logistic-regression-python/?utm_campaign=News&utm_medium=Community&utm_source=DataCamp.com
[Accessed 29 11 2021].

Stoll, A., Ziegele, M. & Quiring, O., 2020. Detecting Impoliteness and Incivility in Online Discussions Classification Approaches for German User Comments. *Computational Communication Research,* 2(1), pp. 109-134.

Subasi, A., 2020. *Practical Machine Learning for Data Analysis Using Python.* 1st ed. s.l.:Elsevier.

Tan, L., 2015. Code Comment Analysis for Improving Software Quality. In: *The Art and Science of Analyzing Software Data.* s.l.:Elsevier, pp. 493-517.

Uysal, A. K. & Gunal, S., 2014. The impact of preprocessing on text classification. *Information Processing and Management,* 50(1), pp. 104-112.

Villatoro-Tello, E. et al., 2012. *A Two-step Approach for Effective Detection of Misbehaving Users in Chats.* Rome, s.n.

Vizer, L. M., Zhou, L. & Sears, A., 2009. Automated Stress Detection Using Keystroke and Linguistic Features: An Exploratory Study. *International Journal of Human-Computer Studies,* 67(10), pp. 870-886.

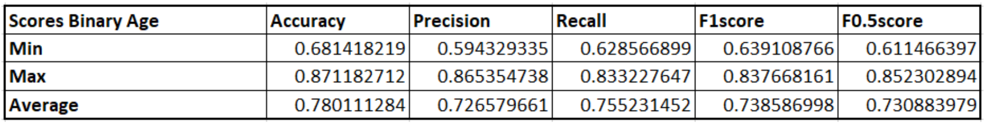
Vrij, A., Fisher, R., Mann, S. & Leal, S., 2006. Detecting Deception by Manipulating Cognitive Load. *Trends in Cognitive Sciences,* 10(4), pp. 141-142.

A. Detailed results

A.1. Results of the binary classifiers

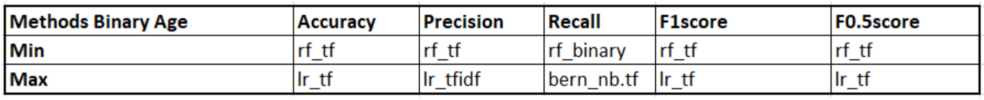
A.1.1. Age classifier

A.1.1.1. Method I (soft vote)



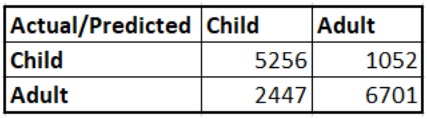
Scores Binary Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.681418219	0.594329335	0.628566899	0.639108766	0.611466397
Max	0.871182712	0.865354738	0.833227647	0.837668161	0.852302894
Average	0.780111284	0.726579661	0.755231452	0.738586998	0.730883979

Appendix table 1 The best and worst results from binary age classification using soft voting



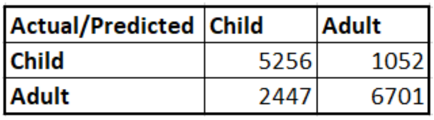
Methods Binary Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tf	rf_tf	rf_binary	rf_tf	rf_tf
Max	lr_tf	lr_tfidf	bern_nb.tf	lr_tf	lr_tf

Appendix table 2 The methods corresponding to the best and worst results from binary age classification using soft voting



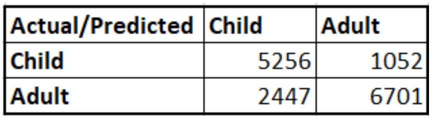
Actual/Predicted	Child	Adult
Child	5256	1052
Adult	2447	6701

Appendix table 3 Bernoulli NB TF on binary age data soft vote



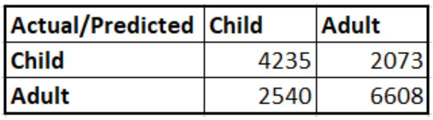
Actual/Predicted	Child	Adult
Child	5256	1052
Adult	2447	6701

Appendix table 4 Bernoulli NB Binary on binary age data soft vote



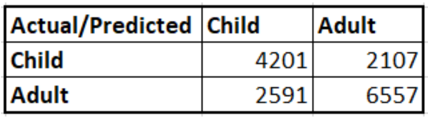
Actual/Predicted	Child	Adult
Child	5256	1052
Adult	2447	6701

Appendix table 5 Bernoulli NB TF-IDF on binary age data soft vote



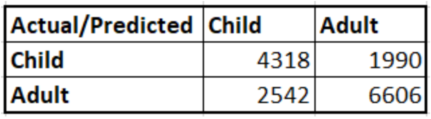
Actual/Predicted	Child	Adult
Child	4235	2073
Adult	2540	6608

Appendix table 6 DT Binary on binary age data soft vote



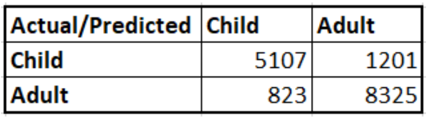
Actual/Predicted	Child	Adult
Child	4201	2107
Adult	2591	6557

Appendix table 7 DT TF on binary age data soft vote



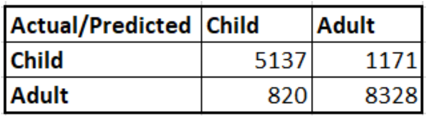
Actual/Predicted	Child	Adult
Child	4318	1990
Adult	2542	6606

Appendix table 8 DT TF-IDF on binary age data soft vote



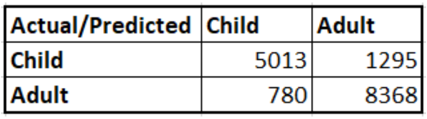
Actual/Predicted	Child	Adult
Child	5107	1201
Adult	823	8325

Appendix table 9 LR Binary on binary age data soft vote



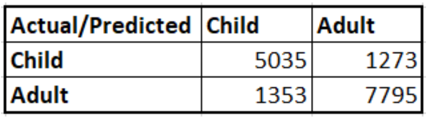
Actual/Predicted	Child	Adult
Child	5137	1171
Adult	820	8328

Appendix table 10 LR TF on binary age data soft vote



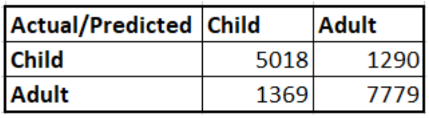
Actual/Predicted	Child	Adult
Child	5013	1295
Adult	780	8368

Appendix table 11 LR TF-IDF on binary age data soft vote



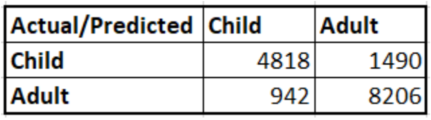
Actual/Predicted	Child	Adult
Child	5035	1273
Adult	1353	7795

Appendix table 12 Multinomial NB Binary on binary age data soft vote



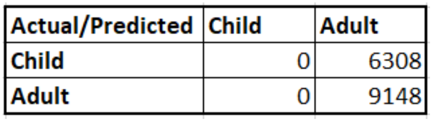
Actual/Predicted	Child	Adult
Child	5018	1290
Adult	1369	7779

Appendix table 13 Multinomial NB TF on binary age data soft vote



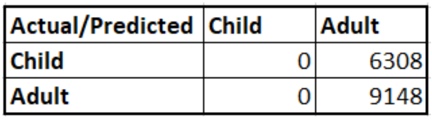
Actual/Predicted	Child	Adult
Child	4818	1490
Adult	942	8206

Appendix table 14 Multinomial NB TF-IDF on binary age data soft vote



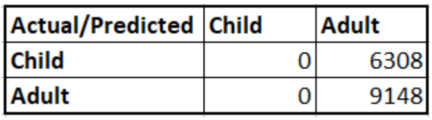
Actual/Predicted	Child	Adult
Child	0	6308
Adult	0	9148

Appendix table 15 NN Binary on binary age data soft vote



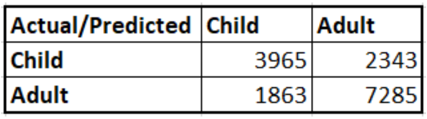
Actual/Predicted	Child	Adult
Child	0	6308
Adult	0	9148

Appendix table 16 NN TF on binary age data soft vote



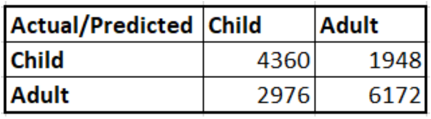
Actual/Predicted	Child	Adult
Child	0	6308
Adult	0	9148

Appendix table 17 NN TF-IDF on binary age data soft vote



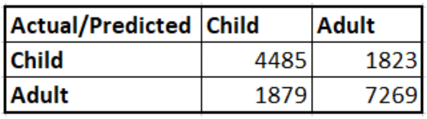
Actual/Predicted	Child	Adult
Child	3965	2343
Adult	1863	7285

Appendix table 18 RF Binary on binary age data soft vote



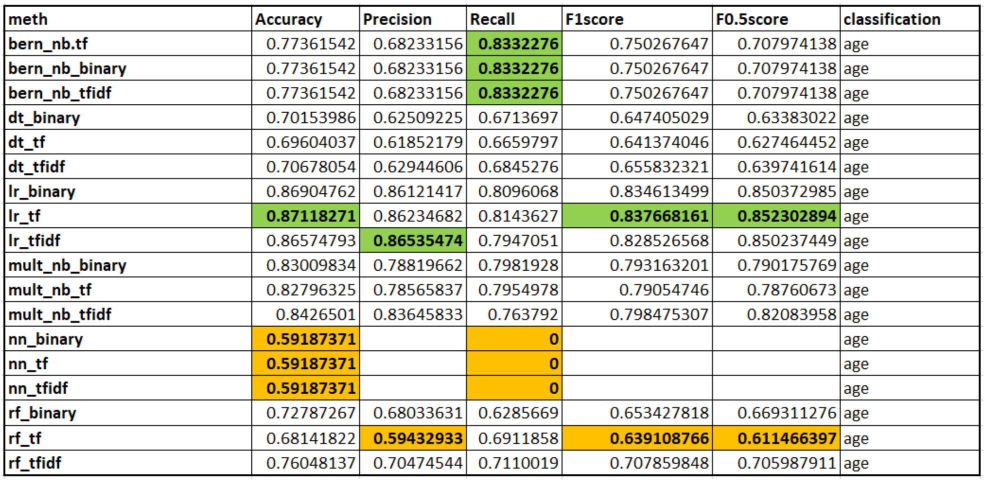
Actual/Predicted	Child	Adult
Child	4360	1948
Adult	2976	6172

Appendix table 19 RF TF on binary age data soft vote



Actual/Predicted	Child	Adult
Child	4485	1823
Adult	1879	7269

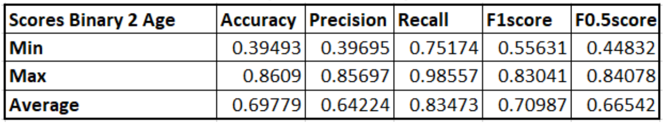
Appendix table 20 RF TF-IDF on binary age data soft vote



meth	Accuracy	Precision	Recall	F1score	F0.5score	classification
bern_nb.tf	0.77361542	0.68233156	**0.8332276**	0.750267647	0.707974138	age
bern_nb_binary	0.77361542	0.68233156	**0.8332276**	0.750267647	0.707974138	age
bern_nb_tfidf	0.77361542	0.68233156	**0.8332276**	0.750267647	0.707974138	age
dt_binary	0.70153986	0.62509225	0.6713697	0.647405029	0.63383022	age
dt_tf	0.69604037	0.61852179	0.6659797	0.641374046	0.627464452	age
dt_tfidf	0.70678054	0.62944606	0.6845276	0.655832321	0.639741614	age
lr_binary	0.86904762	0.86121417	0.8096068	0.834613499	0.850372985	age
lr_tf	**0.87118271**	0.86234682	0.8143627	**0.837668161**	**0.852302894**	age
lr_tfidf	0.86574793	**0.86535474**	0.7947051	0.828526568	0.850237449	age
mult_nb_binary	0.83009834	0.78819662	0.7981928	0.793163201	0.790175769	age
mult_nb_tf	0.82796325	0.78565837	0.7954978	0.79054746	0.78760673	age
mult_nb_tfidf	0.8426501	0.83645833	0.763792	0.798475307	0.82083958	age
nn_binary	**0.59187371**		**0**			age
nn_tf	**0.59187371**		**0**			age
nn_tfidf	**0.59187371**		**0**			age
rf_binary	0.72787267	0.68033631	0.6285669	0.653427818	0.669311276	age
rf_tf	0.68141822	**0.59432933**	0.6911858	**0.639108766**	**0.611466397**	age
rf_tfidf	0.76048137	0.70474544	0.7110019	0.707859848	0.705987911	age

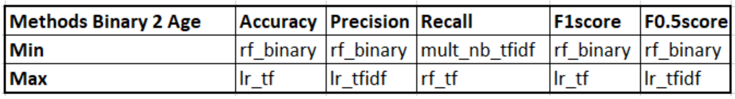
Appendix table 21 All metrics and methods for binary age classification soft vote

A.1.1.2. Method II (hard vote)



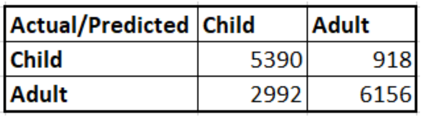
Scores Binary 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.39493	0.39695	0.75174	0.55631	0.44832
Max	0.8609	0.85697	0.98557	0.83041	0.84078
Average	0.69779	0.64224	0.83473	0.70987	0.66542

Appendix table 22 The best and worst results from binary age classification using hard voting



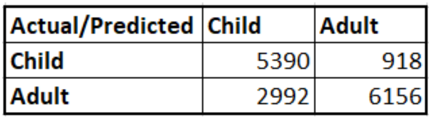
Methods Binary 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_binary	rf_binary	mult_nb_tfidf	rf_binary	rf_binary
Max	lr_tf	lr_tfidf	rf_tf	lr_tf	lr_tfidf

Appendix table 23 The methods corresponding to the best and worst results from binary age classification using hard voting



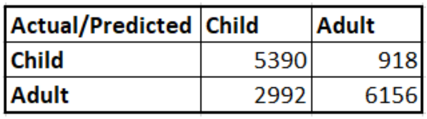
Actual/Predicted	Child	Adult
Child	5390	918
Adult	2992	6156

Appendix table 24 Bernoulli NB TF on binary age data hard vote



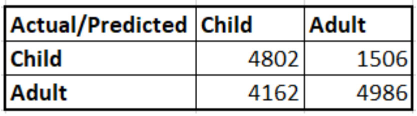
Actual/Predicted	Child	Adult
Child	5390	918
Adult	2992	6156

Appendix table 25 Bernoulli NB Binary on binary age data hard vote



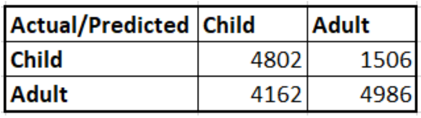
Actual/Predicted	Child	Adult
Child	5390	918
Adult	2992	6156

Appendix table 26 Bernoulli NB TF-IDF on binary age data hard vote



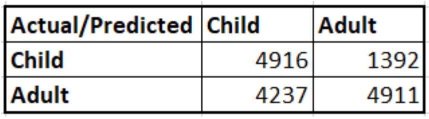
Actual/Predicted	Child	Adult
Child	4802	1506
Adult	4162	4986

Appendix table 27 DT Binary on binary age data hard vote



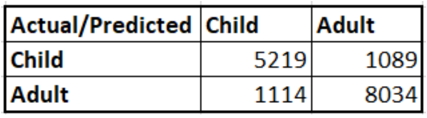
Actual/Predicted	Child	Adult
Child	4802	1506
Adult	4162	4986

Appendix table 28 DT TF on binary age data hard vote



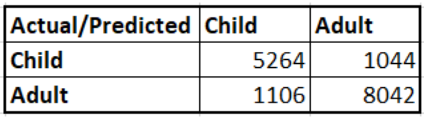
Actual/Predicted	Child	Adult
Child	4916	1392
Adult	4237	4911

Appendix table 29 DT TF-IDF on binary age data hard vote



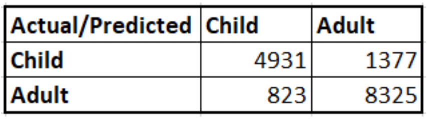
Actual/Predicted	Child	Adult
Child	5219	1089
Adult	1114	8034

Appendix table 30 LR Binary on binary age data hard vote



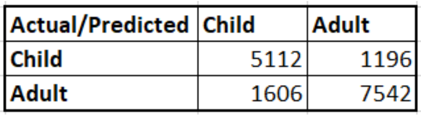
Actual/Predicted	Child	Adult
Child	5264	1044
Adult	1106	8042

Appendix table 31 LR TF on binary age data hard vote



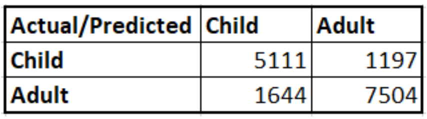
Actual/Predicted	Child	Adult
Child	4931	1377
Adult	823	8325

Appendix table 32 LR TF-IDF on binary age data hard vote



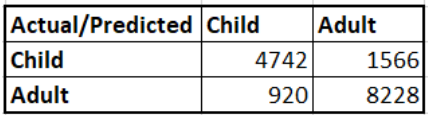
Actual/Predicted	Child	Adult
Child	5112	1196
Adult	1606	7542

Appendix table 33 Multinomial NB Binary on binary age data hard vote



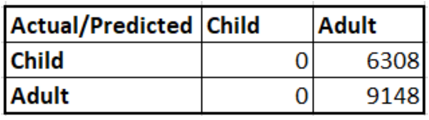
Actual/Predicted	Child	Adult
Child	5111	1197
Adult	1644	7504

Appendix table 34 Multinomial NB TF on binary age data hard vote



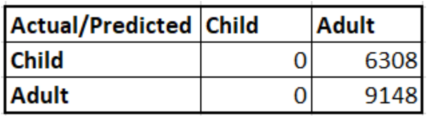
Actual/Predicted	Child	Adult
Child	4742	1566
Adult	920	8228

Appendix table 35 Multinomial NB TF-IDF on binary age data hard vote



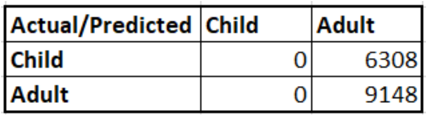
Actual/Predicted	Child	Adult
Child	0	6308
Adult	0	9148

Appendix table 36 NN Binary on binary age data hard vote



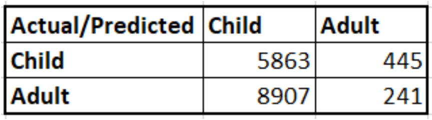
Actual/Predicted	Child	Adult
Child	0	6308
Adult	0	9148

Appendix table 37 NN TF on binary age data hard vote



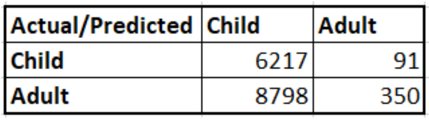
Actual/Predicted	Child	Adult
Child	0	6308
Adult	0	9148

Appendix table 38 NN TF-IDF on binary age data hard vote



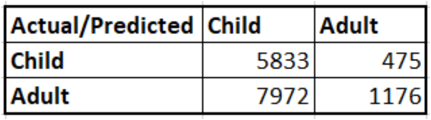
Actual/Predicted	Child	Adult
Child	5863	445
Adult	8907	241

Appendix table 39 RF Binary on binary age data hard vote



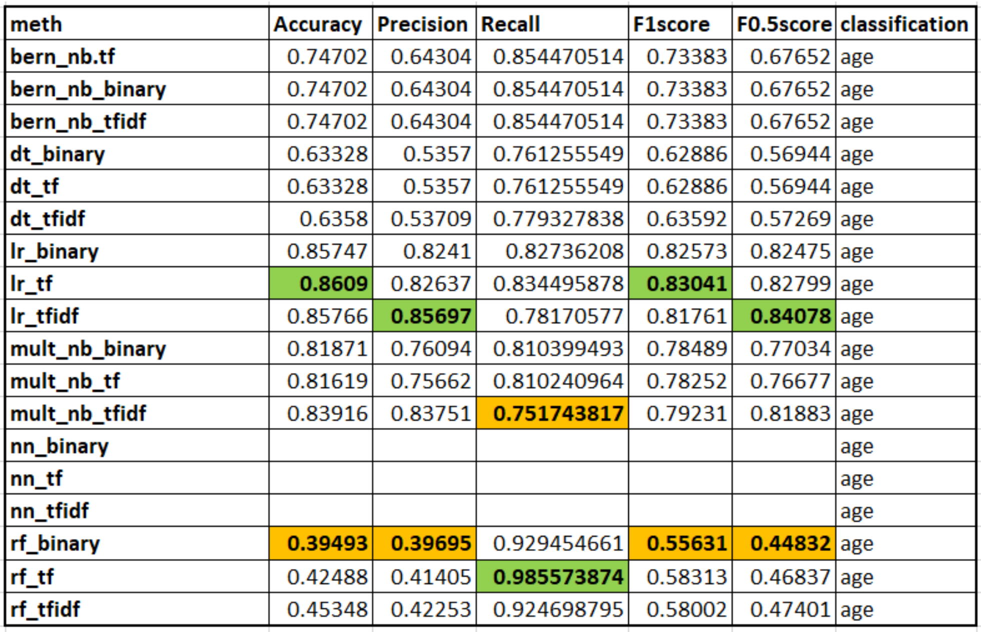
Actual/Predicted	Child	Adult
Child	6217	91
Adult	8798	350

Appendix table 40 RF TF on binary age data hard vote



Actual/Predicted	Child	Adult
Child	5833	475
Adult	7972	1176

Appendix table 41 RF TF-IDF on binary age data hard vote

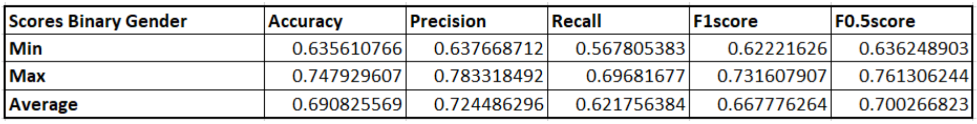


meth	Accuracy	Precision	Recall	F1score	F0.5score	classification
bern_nb.tf	0.74702	0.64304	0.854470514	0.73383	0.67652	age
bern_nb_binary	0.74702	0.64304	0.854470514	0.73383	0.67652	age
bern_nb_tfidf	0.74702	0.64304	0.854470514	0.73383	0.67652	age
dt_binary	0.63328	0.5357	0.761255549	0.62886	0.56944	age
dt_tf	0.63328	0.5357	0.761255549	0.62886	0.56944	age
dt_tfidf	0.6358	0.53709	0.779327838	0.63592	0.57269	age
lr_binary	0.85747	0.8241	0.82736208	0.82573	0.82475	age
lr_tf	**0.8609**	0.82637	0.834495878	**0.83041**	0.82799	age
lr_tfidf	0.85766	**0.85697**	0.78170577	0.81761	**0.84078**	age
mult_nb_binary	0.81871	0.76094	0.810399493	0.78489	0.77034	age
mult_nb_tf	0.81619	0.75662	0.810240964	0.78252	0.76677	age
mult_nb_tfidf	0.83916	0.83751	**0.751743817**	0.79231	0.81883	age
nn_binary						age
nn_tf						age
nn_tfidf						age
rf_binary	**0.39493**	**0.39695**	0.929454661	**0.55631**	**0.44832**	age
rf_tf	0.42488	0.41405	**0.985573874**	0.58313	0.46837	age
rf_tfidf	0.45348	0.42253	0.924698795	0.58002	0.47401	age

Appendix table 42 All metrics and methods for binary age classification hard vote

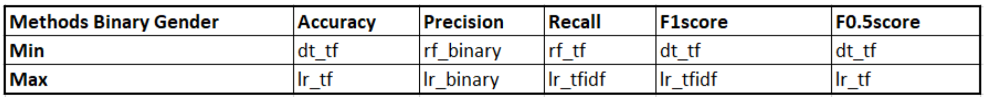
A.1.2. Gender classifier

A.1.2.1. Method I (soft vote)



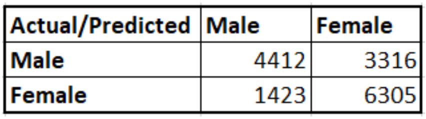
Scores Binary Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.635610766	0.637668712	0.567805383	0.62221626	0.636248903
Max	0.747929607	0.783318492	0.69681677	0.731607907	0.761306244
Average	0.690825569	0.724486296	0.621756384	0.667776264	0.700266823

Appendix table 43 The best and worst results from binary gender classification using soft voting



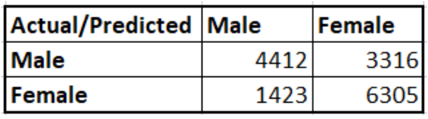
Methods Binary Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	dt_tf	rf_binary	rf_tf	dt_tf	dt_tf
Max	lr_tf	lr_binary	lr_tfidf	lr_tfidf	lr_tf

Appendix table 44 The methods corresponding to the best and worst results from binary gender classification using soft voting



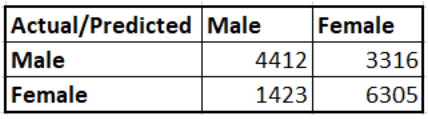
Actual/Predicted	Male	Female
Male	4412	3316
Female	1423	6305

Appendix table 45 Bernoulli NB TF on binary gender data soft vote



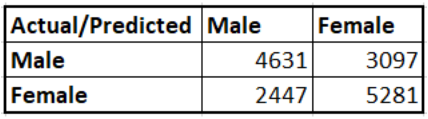
Actual/Predicted	Male	Female
Male	4412	3316
Female	1423	6305

Appendix table 46 Bernoulli NB Binary on binary gender data soft vote



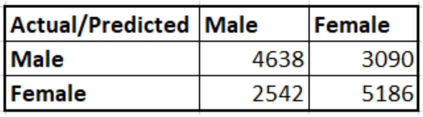
Actual/Predicted	Male	Female
Male	4412	3316
Female	1423	6305

Appendix table 47 Bernoulli NB TF-IDF on binary gender data soft vote



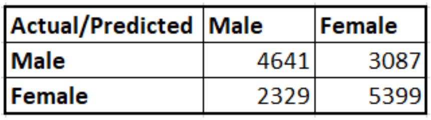
Actual/Predicted	Male	Female
Male	4631	3097
Female	2447	5281

Appendix table 48 DT Binary on binary gender data soft vote



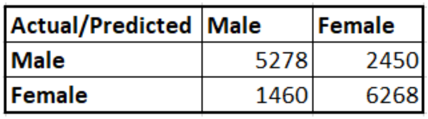
Actual/Predicted	Male	Female
Male	4638	3090
Female	2542	5186

Appendix table 49 DT TF on binary gender data soft vote



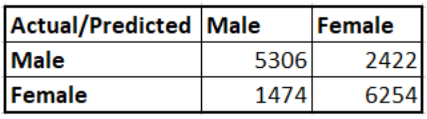
Actual/Predicted	Male	Female
Male	4641	3087
Female	2329	5399

Appendix table 50 DT TF-IDF on binary gender data soft vote



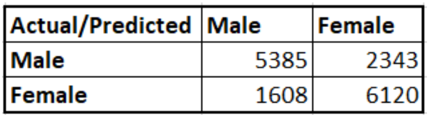
Actual/Predicted	Male	Female
Male	5278	2450
Female	1460	6268

Appendix table 51 LR Binary on binary gender data soft vote



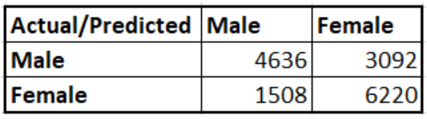
Actual/Predicted	Male	Female
Male	5306	2422
Female	1474	6254

Appendix table 52 LR TF on binary gender data soft vote



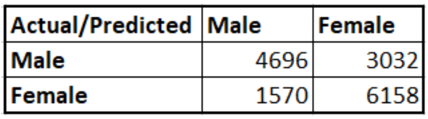
Actual/Predicted	Male	Female
Male	5385	2343
Female	1608	6120

Appendix table 53 LR TF-IDF on binary gender data soft vote



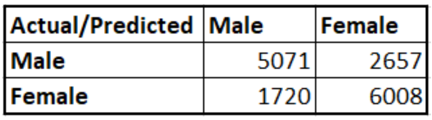
Actual/Predicted	Male	Female
Male	4636	3092
Female	1508	6220

Appendix table 54 Multinomial NB Binary on binary gender data soft vote



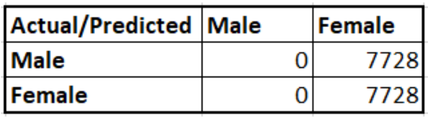
Actual/Predicted	Male	Female
Male	4696	3032
Female	1570	6158

Appendix table 55 Multinomial NB TF on binary gender data soft vote



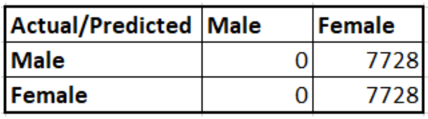
Actual/Predicted	Male	Female
Male	5071	2657
Female	1720	6008

Appendix table 56 Multinomial NB TF-IDF on binary gender data soft vote



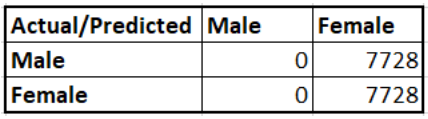
Actual/Predicted	Male	Female
Male	0	7728
Female	0	7728

Appendix table 57 NN Binary on binary gender data soft vote



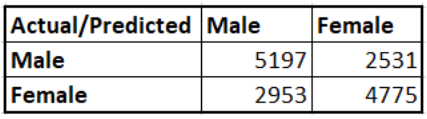
Actual/Predicted	Male	Female
Male	0	7728
Female	0	7728

Appendix table 58 NN TF on binary gender data soft vote



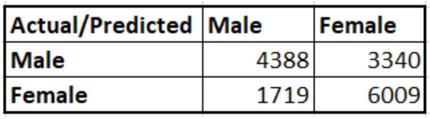
Actual/Predicted	Male	Female
Male	0	7728
Female	0	7728

Appendix table 59 NN TF-IDF on binary gender data soft vote



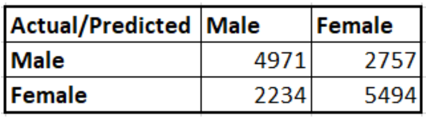
Actual/Predicted	Male	Female
Male	5197	2531
Female	2953	4775

Appendix table 60 RF Binary on binary gender data soft vote



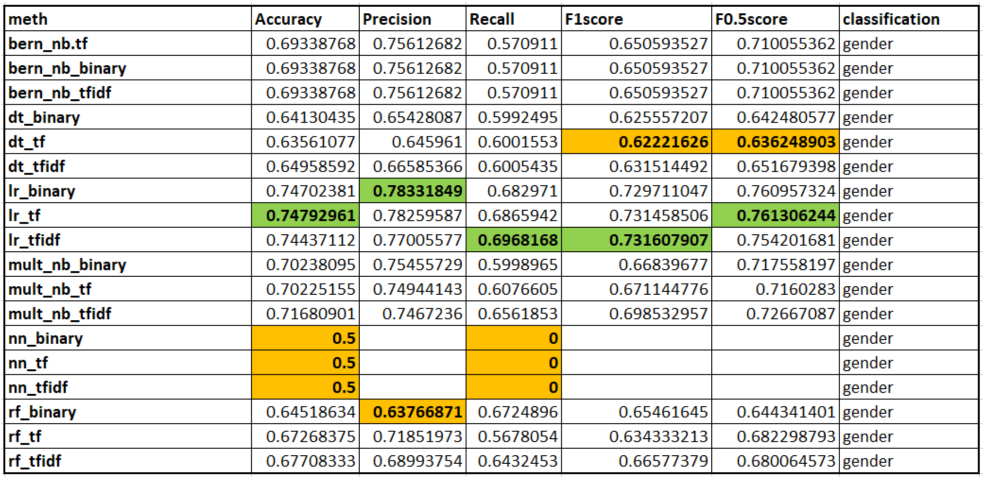
Actual/Predicted	Male	Female
Male	4388	3340
Female	1719	6009

Appendix table 61 RF TF on binary gender data soft vote



Actual/Predicted	Male	Female
Male	4971	2757
Female	2234	5494

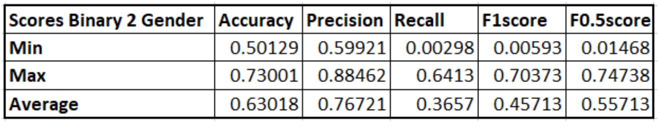
Appendix table 62 RF TF-IDF on binary gender data soft vote



meth	Accuracy	Precision	Recall	F1score	F0.5score	classification
bern_nb.tf	0.69338768	0.75612682	0.570911	0.650593527	0.710055362	gender
bern_nb_binary	0.69338768	0.75612682	0.570911	0.650593527	0.710055362	gender
bern_nb_tfidf	0.69338768	0.75612682	0.570911	0.650593527	0.710055362	gender
dt_binary	0.64130435	0.65428087	0.5992495	0.625557207	0.642480577	gender
dt_tf	0.63561077	0.645961	0.6001553	**0.62221626**	**0.636248903**	gender
dt_tfidf	0.64958592	0.66585366	0.6005435	0.631514492	0.651679398	gender
lr_binary	0.74702381	**0.78331849**	0.682971	0.729711047	0.760957324	gender
lr_tf	**0.74792961**	0.78259587	0.6865942	0.731458506	**0.761306244**	gender
lr_tfidf	0.74437112	0.77005577	**0.6968168**	**0.731607907**	0.754201681	gender
mult_nb_binary	0.70238095	0.75455729	0.5998965	0.66839677	0.717558197	gender
mult_nb_tf	0.70225155	0.74944143	0.6076605	0.671144776	0.7160283	gender
mult_nb_tfidf	0.71680901	0.7467236	0.6561853	0.698532957	0.72667087	gender
nn_binary	**0.5**		**0**			gender
nn_tf	**0.5**		**0**			gender
nn_tfidf	**0.5**		**0**			gender
rf_binary	0.64518634	**0.63766871**	0.6724896	0.65461645	0.644341401	gender
rf_tf	0.67268375	0.71851973	0.5678054	0.634333213	0.682298793	gender
rf_tfidf	0.67708333	0.68993754	0.6432453	0.66577379	0.680064573	gender

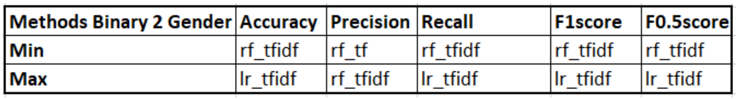
Appendix table 63 All metrics and methods for binary gender classification soft vote

A.1.2.2. Method II (hard vote)



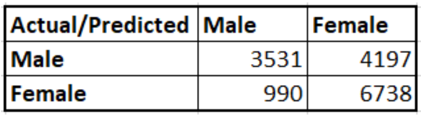
Scores Binary 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.50129	0.59921	0.00298	0.00593	0.01468
Max	0.73001	0.88462	0.6413	0.70373	0.74738
Average	0.63018	0.76721	0.3657	0.45713	0.55713

Appendix table 64 The best and worst results from binary gender classification using hard voting



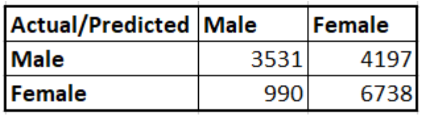
Methods Binary 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tfidf	rf_tf	rf_tfidf	rf_tfidf	rf_tfidf
Max	lr_tfidf	rf_tfidf	lr_tfidf	lr_tfidf	lr_tfidf

Appendix table 65 The methods corresponding to the best and worst results from binary gender classification using hard voting



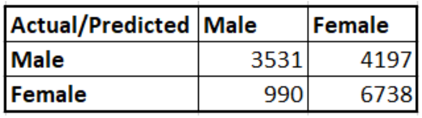
Actual/Predicted	Male	Female
Male	3531	4197
Female	990	6738

Appendix table 66 Bernoulli NB TF on binary gender data hard vote



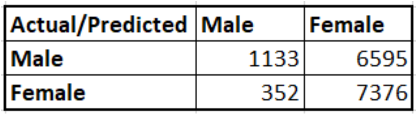
Actual/Predicted	Male	Female
Male	3531	4197
Female	990	6738

Appendix table 67 Bernoulli NB Binary on binary gender data hard vote



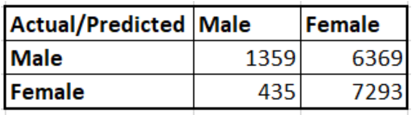
Actual/Predicted	Male	Female
Male	3531	4197
Female	990	6738

Appendix table 68 Bernoulli NB TF-IDF on binary gender data hard vote



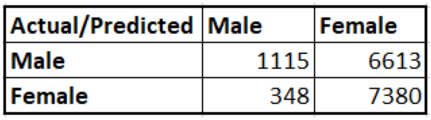
Actual/Predicted	Male	Female
Male	1133	6595
Female	352	7376

Appendix table 69 DT Binary on binary gender data hard vote



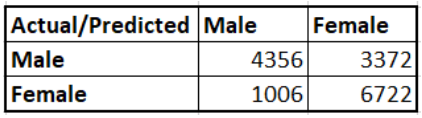
Actual/Predicted	Male	Female
Male	1359	6369
Female	435	7293

Appendix table 70 DT TF on binary gender data hard vote



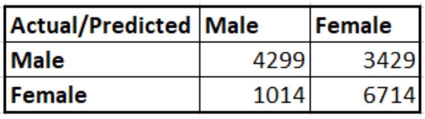
Actual/Predicted	Male	Female
Male	1115	6613
Female	348	7380

Appendix table 71 DT TF-IDF on binary gender data hard vote



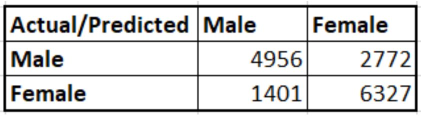
Actual/Predicted	Male	Female
Male	4356	3372
Female	1006	6722

Appendix table 72 LR Binary on binary gender data hard vote



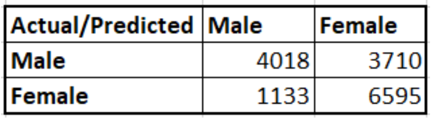
Actual/Predicted	Male	Female
Male	4299	3429
Female	1014	6714

Appendix table 73 LR TF on binary gender data hard vote



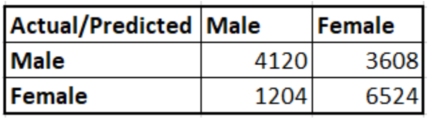
Actual/Predicted	Male	Female
Male	4956	2772
Female	1401	6327

Appendix table 74 LR TF-IDF on binary gender data hard vote



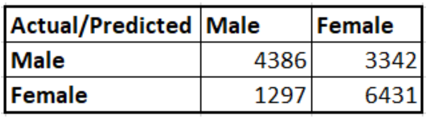
Actual/Predicted	Male	Female
Male	4018	3710
Female	1133	6595

Appendix table 75 Multinomial NB Binary on binary gender data hard vote



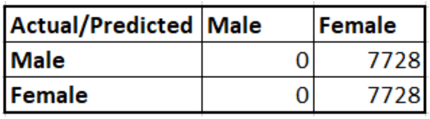
Actual/Predicted	Male	Female
Male	4120	3608
Female	1204	6524

Appendix table 76 Multinomial NB TF on binary gender data hard vote



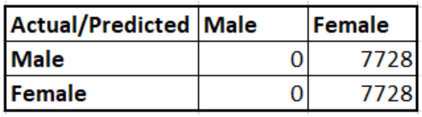
Actual/Predicted	Male	Female
Male	4386	3342
Female	1297	6431

Appendix table 77 Multinomial NB TF-IDF on binary gender data hard vote



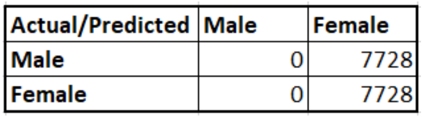
Actual/Predicted	Male	Female
Male	0	7728
Female	0	7728

Appendix table 78 NN Binary on binary gender data hard vote



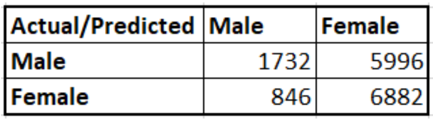
Actual/Predicted	Male	Female
Male	0	7728
Female	0	7728

Appendix table 79 NN TF on binary gender data hard vote



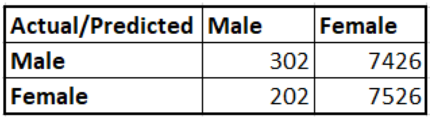
Actual/Predicted	Male	Female
Male	0	7728
Female	0	7728

Appendix table 80 NN TF-IDF on binary gender data hard vote



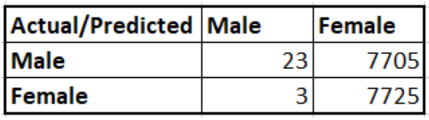
Actual/Predicted	Male	Female
Male	1732	5996
Female	846	6882

Appendix table 81 RF Binary on binary gender data hard vote



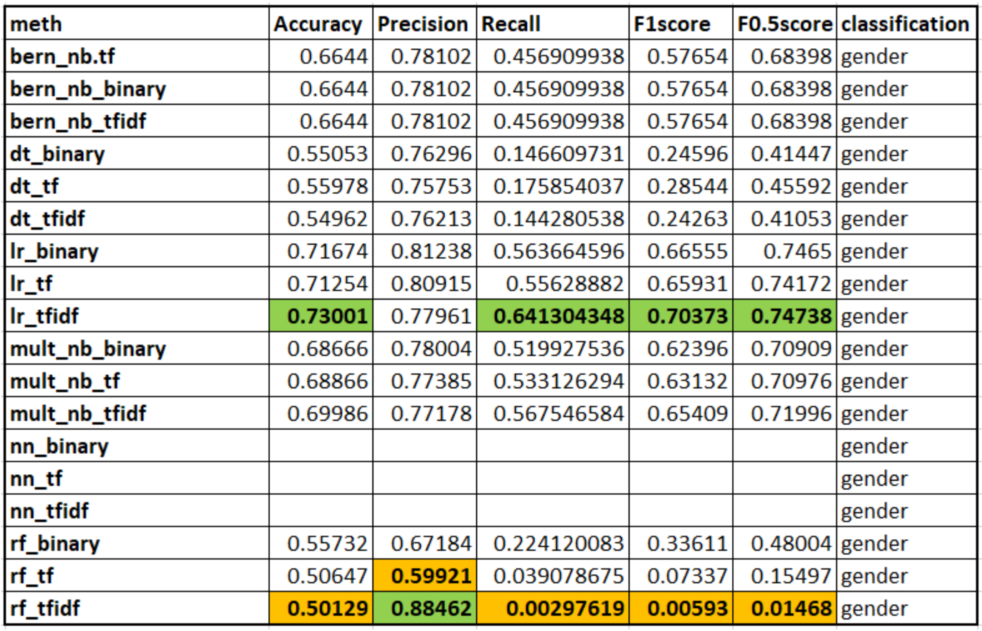
Actual/Predicted	Male	Female
Male	302	7426
Female	202	7526

Appendix table 82 RF TF on binary gender data hard vote



Actual/Predicted	Male	Female
Male	23	7705
Female	3	7725

Appendix table 83 RF TF-IDF on binary gender data hard vote

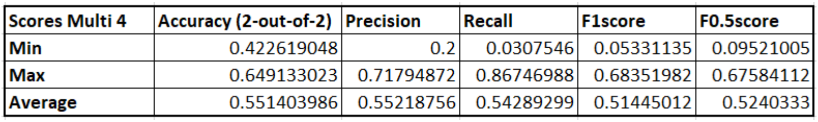


meth	Accuracy	Precision	Recall	F1score	F0.5score	classification
bern_nb.tf	0.6644	0.78102	0.456909938	0.57654	0.68398	gender
bern_nb_binary	0.6644	0.78102	0.456909938	0.57654	0.68398	gender
bern_nb_tfidf	0.6644	0.78102	0.456909938	0.57654	0.68398	gender
dt_binary	0.55053	0.76296	0.146609731	0.24596	0.41447	gender
dt_tf	0.55978	0.75753	0.175854037	0.28544	0.45592	gender
dt_tfidf	0.54962	0.76213	0.144280538	0.24263	0.41053	gender
lr_binary	0.71674	0.81238	0.563664596	0.66555	0.7465	gender
lr_tf	0.71254	0.80915	0.55628882	0.65931	0.74172	gender
lr_tfidf	**0.73001**	0.77961	**0.641304348**	**0.70373**	**0.74738**	gender
mult_nb_binary	0.68666	0.78004	0.519927536	0.62396	0.70909	gender
mult_nb_tf	0.68866	0.77385	0.533126294	0.63132	0.70976	gender
mult_nb_tfidf	0.69986	0.77178	0.567546584	0.65409	0.71996	gender
nn_binary						gender
nn_tf						gender
nn_tfidf						gender
rf_binary	0.55732	0.67184	0.224120083	0.33611	0.48004	gender
rf_tf	0.50647	**0.59921**	0.039078675	0.07337	0.15497	gender
rf_tfidf	**0.50129**	**0.88462**	**0.00297619**	**0.00593**	**0.01468**	gender

Appendix table 84 All metrics and methods for binary gender classification hard vote

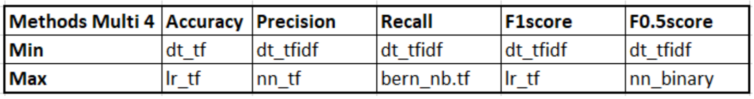
A.2. Results of the 4-class classifiers

A.2.1. Method I (soft vote)



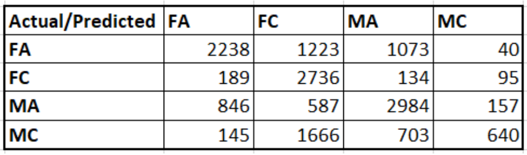
Scores Multi 4	Accuracy (2-out-of-2)	Precision	Recall	F1score	F0.5score
Min	0.422619048	0.2	0.0307546	0.05331135	0.09521005
Max	0.649133023	0.71794872	0.86746988	0.68351982	0.67584112
Average	0.551403986	0.55218756	0.54289299	0.51445012	0.5240333

Appendix table 85 The best and worst results from 4-class classification using soft voting



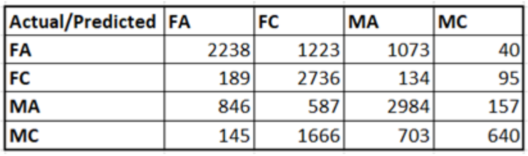
Methods Multi 4	Accuracy	Precision	Recall	F1score	F0.5score
Min	dt_tf	dt_tfidf	dt_tfidf	dt_tfidf	dt_tfidf
Max	lr_tf	nn_tf	bern_nb.tf	lr_tf	nn_binary

Appendix table 86 The methods corresponding to the best and worst results from 4-class classification using soft voting



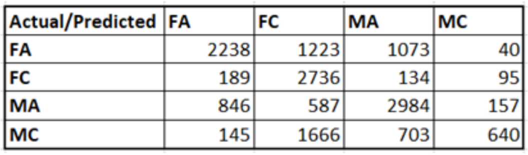
Actual/Predicted	FA	FC	MA	MC
FA	2238	1223	1073	40
FC	189	2736	134	95
MA	846	587	2984	157
MC	145	1666	703	640

Appendix table 87 Bernoulli NB TF on 4-class classification soft vote



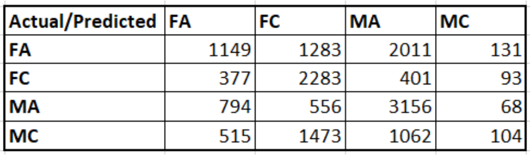
Actual/Predicted	FA	FC	MA	MC
FA	2238	1223	1073	40
FC	189	2736	134	95
MA	846	587	2984	157
MC	145	1666	703	640

Appendix table 88 Bernoulli NB Binary on 4-class classification soft vote



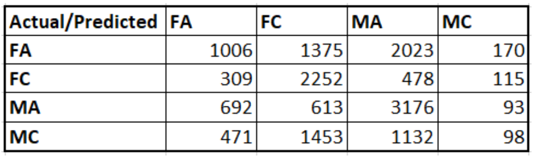
Actual/Predicted	FA	FC	MA	MC
FA	2238	1223	1073	40
FC	189	2736	134	95
MA	846	587	2984	157
MC	145	1666	703	640

Appendix table 89 Bernoulli NB TF-IDF on 4-class classification soft vote



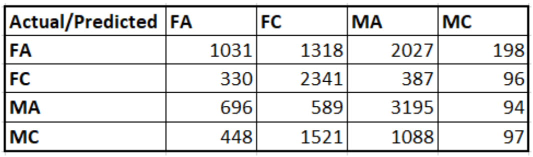
Actual/Predicted	FA	FC	MA	MC
FA	1149	1283	2011	131
FC	377	2283	401	93
MA	794	556	3156	68
MC	515	1473	1062	104

Appendix table 90 DT Binary on 4-class classification soft vote



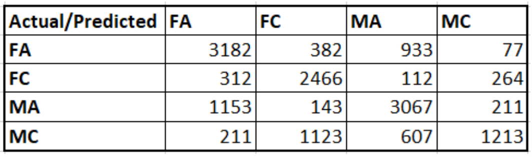
Actual/Predicted	FA	FC	MA	MC
FA	1006	1375	2023	170
FC	309	2252	478	115
MA	692	613	3176	93
MC	471	1453	1132	98

Appendix table 91 DT TF on 4-class classification soft vote



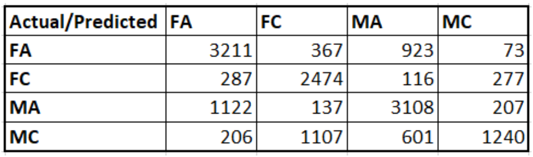
Actual/Predicted	FA	FC	MA	MC
FA	1031	1318	2027	198
FC	330	2341	387	96
MA	696	589	3195	94
MC	448	1521	1088	97

Appendix table 92 DT TF-IDF on 4-class classification soft vote



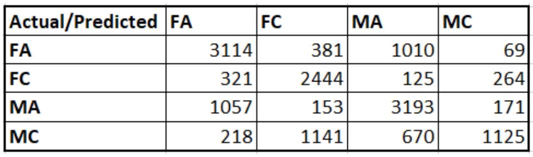
Actual/Predicted	FA	FC	MA	MC
FA	3182	382	933	77
FC	312	2466	112	264
MA	1153	143	3067	211
MC	211	1123	607	1213

Appendix table 93 LR Binary on 4-class classification soft vote



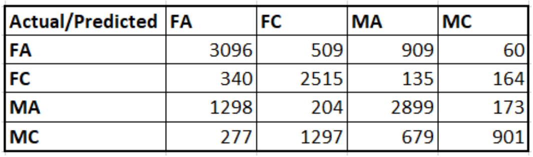
Actual/Predicted	FA	FC	MA	MC
FA	3211	367	923	73
FC	287	2474	116	277
MA	1122	137	3108	207
MC	206	1107	601	1240

Appendix table 94 LR TF on 4-class classification soft vote



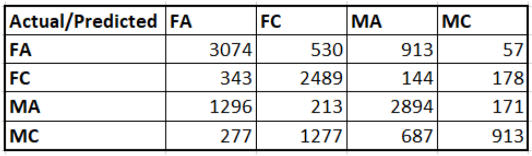
Actual/Predicted	FA	FC	MA	MC
FA	3114	381	1010	69
FC	321	2444	125	264
MA	1057	153	3193	171
MC	218	1141	670	1125

Appendix table 95 LR TF-IDF on 4-class classification soft vote



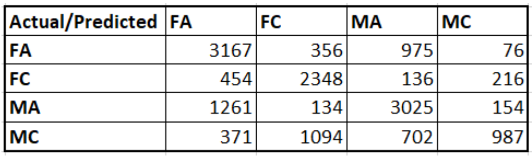
Actual/Predicted	FA	FC	MA	MC
FA	3096	509	909	60
FC	340	2515	135	164
MA	1298	204	2899	173
MC	277	1297	679	901

Appendix table 96 Multinomial NB Binary on 4-class classification soft vote



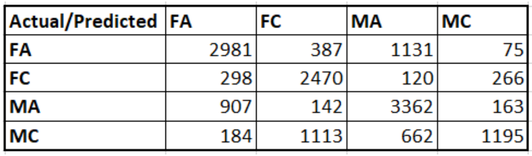
Actual/Predicted	FA	FC	MA	MC
FA	3074	530	913	57
FC	343	2489	144	178
MA	1296	213	2894	171
MC	277	1277	687	913

Appendix table 97 Multinomial NB TF on 4-class classification soft vote



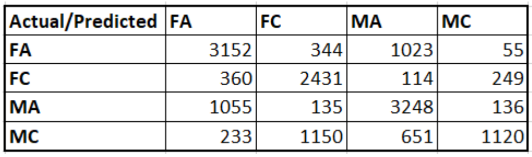
Actual/Predicted	FA	FC	MA	MC
FA	3167	356	975	76
FC	454	2348	136	216
MA	1261	134	3025	154
MC	371	1094	702	987

Appendix table 98 Multinomial NB TF-IDF on 4-class classification soft vote



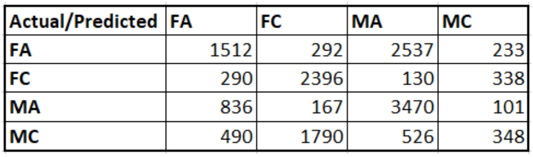
Actual/Predicted	FA	FC	MA	MC
FA	2981	387	1131	75
FC	298	2470	120	266
MA	907	142	3362	163
MC	184	1113	662	1195

Appendix table 99 NN Binary on 4-class classification soft vote



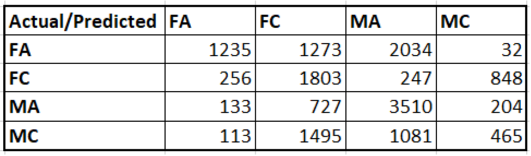
Actual/Predicted	FA	FC	MA	MC
FA	3152	344	1023	55
FC	360	2431	114	249
MA	1055	135	3248	136
MC	233	1150	651	1120

Appendix table 100 NN TF on 4-class classification soft vote



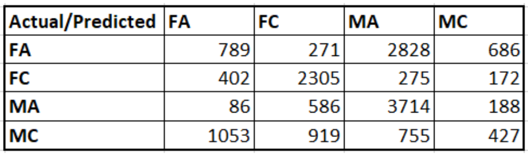
Actual/Predicted	FA	FC	MA	MC
FA	1512	292	2537	233
FC	290	2396	130	338
MA	836	167	3470	101
MC	490	1790	526	348

Appendix table 101 NN TF-IDF on 4-class classification soft vote



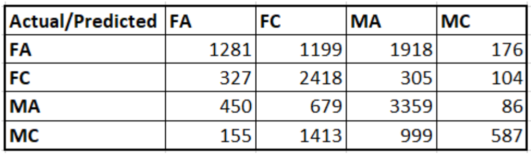
Actual/Predicted	FA	FC	MA	MC
FA	1235	1273	2034	32
FC	256	1803	247	848
MA	133	727	3510	204
MC	113	1495	1081	465

Appendix table 102 RF Binary on 4-class classification soft vote



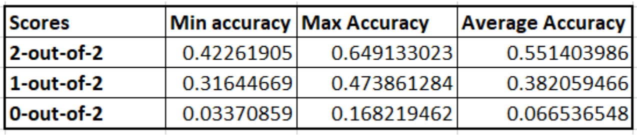
Actual/Predicted	FA	FC	MA	MC
FA	789	271	2828	686
FC	402	2305	275	172
MA	86	586	3714	188
MC	1053	919	755	427

Appendix table 103 RF TF on 4-class classification soft vote



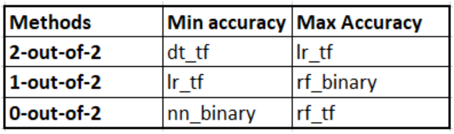
Actual/Predicted	FA	FC	MA	MC
FA	1281	1199	1918	176
FC	327	2418	305	104
MA	450	679	3359	86
MC	155	1413	999	587

Appendix table 104 RF TF-IDF on 4-class classification soft vote



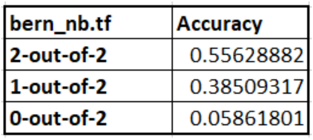
Scores	Min accuracy	Max Accuracy	Average Accuracy
2-out-of-2	0.42261905	0.649133023	0.551403986
1-out-of-2	0.31644669	0.473861284	0.382059466
0-out-of-2	0.03370859	0.168219462	0.066536548

Appendix table 105 An overview of the best and worst accuracies from 4-class classification using soft voting



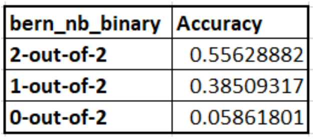
Methods	Min accuracy	Max Accuracy
2-out-of-2	dt_tf	lr_tf
1-out-of-2	lr_tf	rf_binary
0-out-of-2	nn_binary	rf_tf

Appendix table 106 The methods corresponding to the best and worst results from 4-class classification using soft voting



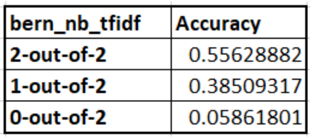
bern_nb.tf	Accuracy
2-out-of-2	0.55628882
1-out-of-2	0.38509317
0-out-of-2	0.05861801

Appendix table 107 Bernoulli NB TF accuracies 4-class classification soft voting



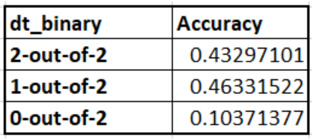
bern_nb_binary	Accuracy
2-out-of-2	0.55628882
1-out-of-2	0.38509317
0-out-of-2	0.05861801

Appendix table 108 Bernoulli NB Binary accuracies 4-class classification soft voting



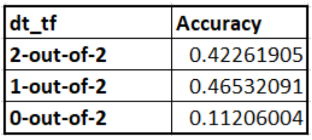
bern_nb_tfidf	Accuracy
2-out-of-2	0.55628882
1-out-of-2	0.38509317
0-out-of-2	0.05861801

Appendix table 109 Bernoulli NB TF-IDF accuracies 4-class classification soft voting



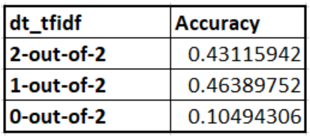
dt_binary	Accuracy
2-out-of-2	0.43297101
1-out-of-2	0.46331522
0-out-of-2	0.10371377

Appendix table 110 DT Binary accuracies 4-class classification soft voting



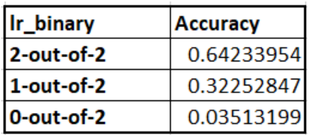
dt_tf	Accuracy
2-out-of-2	0.42261905
1-out-of-2	0.46532091
0-out-of-2	0.11206004

Appendix table 111 DT TF accuracies 4-class classification soft voting



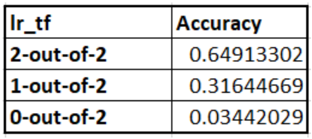
dt_tfidf	Accuracy
2-out-of-2	0.43115942
1-out-of-2	0.46389752
0-out-of-2	0.10494306

Appendix table 112 DT TF-IDF accuracies 4-class classification soft voting



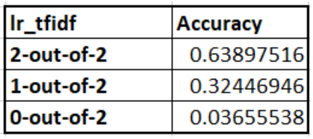
lr_binary	Accuracy
2-out-of-2	0.64233954
1-out-of-2	0.32252847
0-out-of-2	0.03513199

Appendix table 113 LR Binary accuracies 4-class classification soft voting



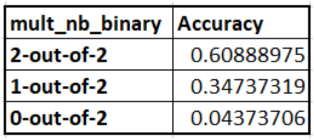
lr_tf	Accuracy
2-out-of-2	0.64913302
1-out-of-2	0.31644669
0-out-of-2	0.03442029

Appendix table 114 LR TF accuracies 4-class classification soft voting



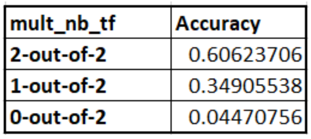
lr_tfidf	Accuracy
2-out-of-2	0.63897516
1-out-of-2	0.32446946
0-out-of-2	0.03655538

Appendix table 115 LR TF-IDF accuracies 4-class classification soft voting



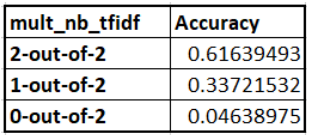
mult_nb_binary	Accuracy
2-out-of-2	0.60888975
1-out-of-2	0.34737319
0-out-of-2	0.04373706

Appendix table 116 Multinomial NB Binary accuracies 4-class classification soft voting



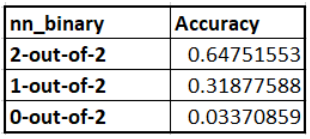
mult_nb_tf	Accuracy
2-out-of-2	0.60623706
1-out-of-2	0.34905538
0-out-of-2	0.04470756

Appendix table 117 Multinomial NB TF accuracies 4-class classification soft voting



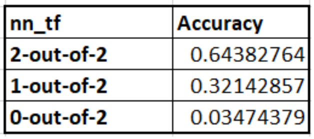
mult_nb_tfidf	Accuracy
2-out-of-2	0.61639493
1-out-of-2	0.33721532
0-out-of-2	0.04638975

Appendix table 118 Multinomial NB TF-IDF accuracies 4-class classification soft voting



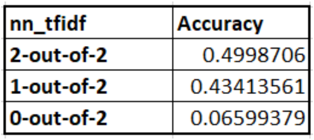
nn_binary	Accuracy
2-out-of-2	0.64751553
1-out-of-2	0.31877588
0-out-of-2	0.03370859

Appendix table 119 NN Binary accuracies 4-class classification soft voting



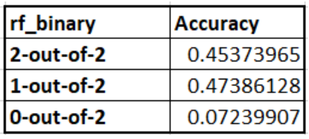
nn_tf	Accuracy
2-out-of-2	0.64382764
1-out-of-2	0.32142857
0-out-of-2	0.03474379

Appendix table 120 NN TF accuracies 4-class classification soft voting



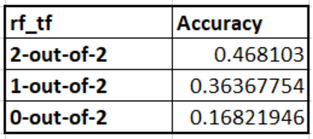
nn_tfidf	Accuracy
2-out-of-2	0.4998706
1-out-of-2	0.43413561
0-out-of-2	0.06599379

Appendix table 121 NN TF-IDF accuracies 4-class classification soft voting



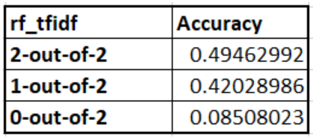
rf_binary	Accuracy
2-out-of-2	0.45373965
1-out-of-2	0.47386128
0-out-of-2	0.07239907

Appendix table 122 RF Binary accuracies 4-class classification soft voting



rf_tf	Accuracy
2-out-of-2	0.468103
1-out-of-2	0.36367754
0-out-of-2	0.16821946

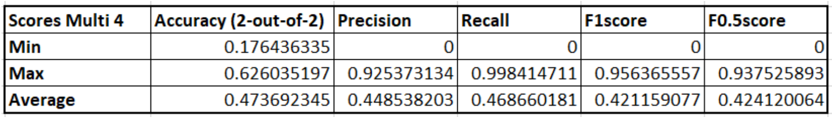
Appendix table 123 RF TF accuracies 4-class classification soft voting



rf_tfidf	Accuracy
2-out-of-2	0.49462992
1-out-of-2	0.42028986
0-out-of-2	0.08508023

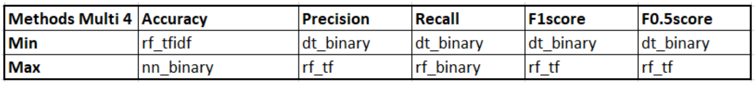
Appendix table 124 RF TF-IDF accuracies 4-class classification soft voting

A.2.2. Method II (hard vote)



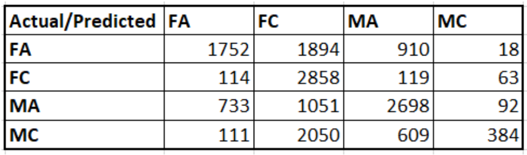
Scores Multi 4	Accuracy (2-out-of-2)	Precision	Recall	F1score	F0.5score
Min	0.176436335	0	0	0	0
Max	0.626035197	0.925373134	0.998414711	0.956365557	0.937525893
Average	0.473692345	0.448538203	0.468660181	0.421159077	0.424120064

Appendix table 125 The best and worst results from 4-class classification using hard voting



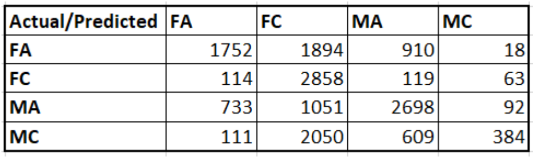
Methods Multi 4	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tfidf	dt_binary	dt_binary	dt_binary	dt_binary
Max	nn_binary	rf_tf	rf_binary	rf_tf	rf_tf

Appendix table 126 The methods corresponding to the best and worst results from 4-class classification using hard voting



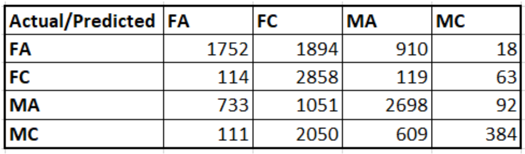
Actual/Predicted	FA	FC	MA	MC
FA	1752	1894	910	18
FC	114	2858	119	63
MA	733	1051	2698	92
MC	111	2050	609	384

Appendix table 127 Bernoulli NB TF on 4-class classification hard vote



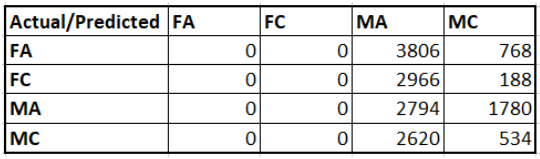
Actual/Predicted	FA	FC	MA	MC
FA	1752	1894	910	18
FC	114	2858	119	63
MA	733	1051	2698	92
MC	111	2050	609	384

Appendix table 128 Bernoulli NB Binary on 4-class classification hard vote



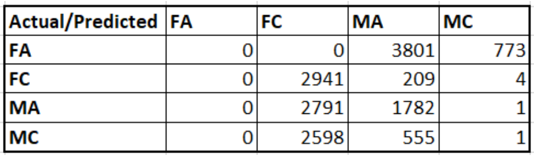
Actual/Predicted	FA	FC	MA	MC
FA	1752	1894	910	18
FC	114	2858	119	63
MA	733	1051	2698	92
MC	111	2050	609	384

Appendix table 129 Bernoulli NB TF-IDF on 4-class classification hard vote



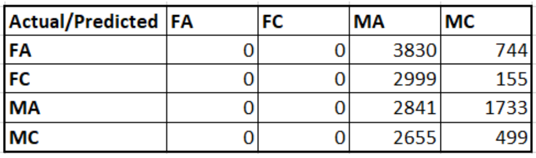
Actual/Predicted	FA	FC	MA	MC
FA	0	0	3806	768
FC	0	0	2966	188
MA	0	0	2794	1780
MC	0	0	2620	534

Appendix table 130 DT Binary on 4-class classification hard vote



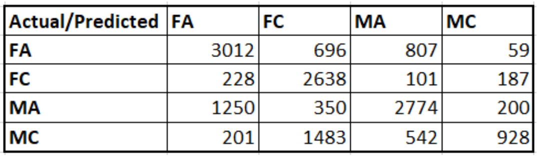
Actual/Predicted	FA	FC	MA	MC
FA	0	0	3801	773
FC	0	2941	209	4
MA	0	2791	1782	1
MC	0	2598	555	1

Appendix table 131 DT TF on 4-class classification hard vote



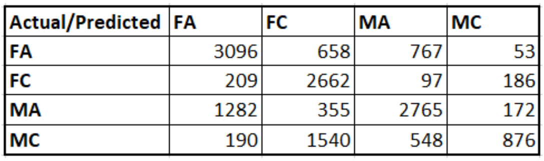
Actual/Predicted	FA	FC	MA	MC
FA	0	0	3830	744
FC	0	0	2999	155
MA	0	0	2841	1733
MC	0	0	2655	499

Appendix table 132 DT TF-IDF on 4-class classification hard vote



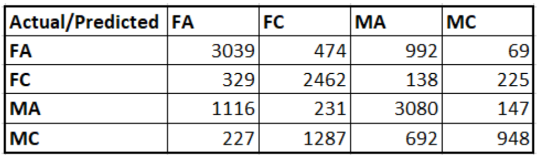
Actual/Predicted	FA	FC	MA	MC
FA	3012	696	807	59
FC	228	2638	101	187
MA	1250	350	2774	200
MC	201	1483	542	928

Appendix table 133 LR Binary on 4-class classification hard vote



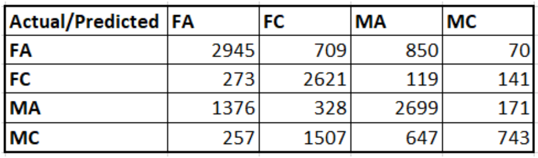
Actual/Predicted	FA	FC	MA	MC
FA	3096	658	767	53
FC	209	2662	97	186
MA	1282	355	2765	172
MC	190	1540	548	876

Appendix table 134 LR TF on 4-class classification hard vote



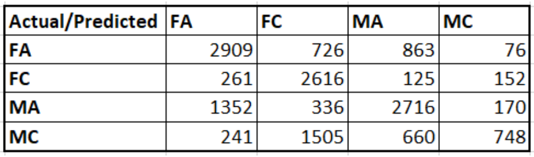
Actual/Predicted	FA	FC	MA	MC
FA	3039	474	992	69
FC	329	2462	138	225
MA	1116	231	3080	147
MC	227	1287	692	948

Appendix table 135 LR TF-IDF on 4-class classification hard vote



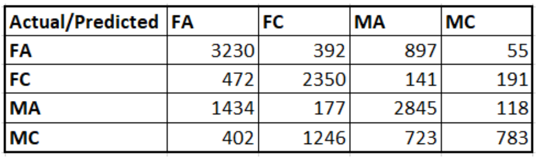
Actual/Predicted	FA	FC	MA	MC
FA	2945	709	850	70
FC	273	2621	119	141
MA	1376	328	2699	171
MC	257	1507	647	743

Appendix table 136 Multinomial NB Binary on 4-class classification hard vote



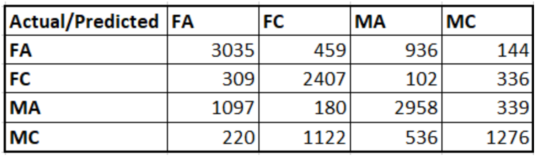
Actual/Predicted	FA	FC	MA	MC
FA	2909	726	863	76
FC	261	2616	125	152
MA	1352	336	2716	170
MC	241	1505	660	748

Appendix table 137 Multinomial NB TF on 4-class classification hard vote



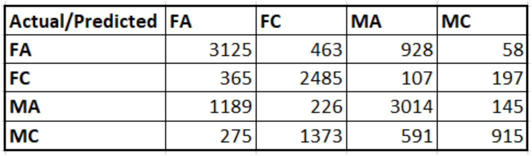
Actual/Predicted	FA	FC	MA	MC
FA	3230	392	897	55
FC	472	2350	141	191
MA	1434	177	2845	118
MC	402	1246	723	783

Appendix table 138 Multinomial NB TF-IDF on 4-class classification hard vote



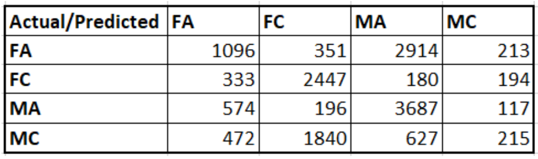
Actual/Predicted	FA	FC	MA	MC
FA	3035	459	936	144
FC	309	2407	102	336
MA	1097	180	2958	339
MC	220	1122	536	1276

Appendix table 139 NN Binary on 4-class classification hard vote



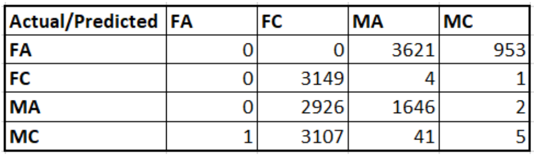
Actual/Predicted	FA	FC	MA	MC
FA	3125	463	928	58
FC	365	2485	107	197
MA	1189	226	3014	145
MC	275	1373	591	915

Appendix table 140 NN TF on 4-class classification hard vote



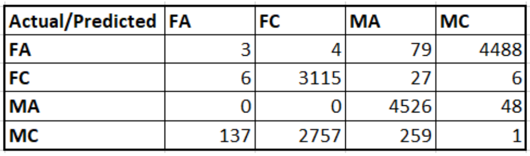
Actual/Predicted	FA	FC	MA	MC
FA	1096	351	2914	213
FC	333	2447	180	194
MA	574	196	3687	117
MC	472	1840	627	215

Appendix table 141 NN TF-IDF on 4-class classification hard vote



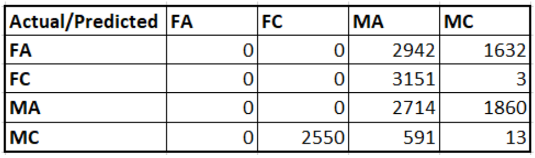
Actual/Predicted	FA	FC	MA	MC
FA	0	0	3621	953
FC	0	3149	4	1
MA	0	2926	1646	2
MC	1	3107	41	5

Appendix table 142 RF Binary on 4-class classification hard vote



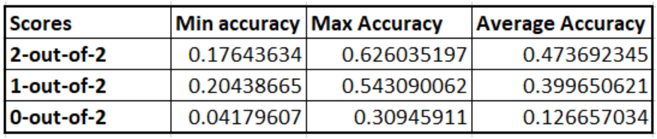
Actual/Predicted	FA	FC	MA	MC
FA	3	4	79	4488
FC	6	3115	27	6
MA	0	0	4526	48
MC	137	2757	259	1

Appendix table 143 RF TF on 4-class classification hard vote



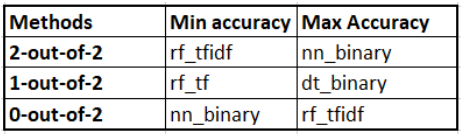
Actual/Predicted	FA	FC	MA	MC
FA	0	0	2942	1632
FC	0	0	3151	3
MA	0	0	2714	1860
MC	0	2550	591	13

Appendix table 144 RF TF-IDF on 4-class classification hard vote



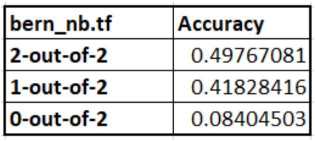
Scores	Min accuracy	Max Accuracy	Average Accuracy
2-out-of-2	0.17643634	0.626035197	0.473692345
1-out-of-2	0.20438665	0.543090062	0.399650621
0-out-of-2	0.04179607	0.30945911	0.126657034

Appendix table 145 An overview of the best and worst accuracies from 4-class classification using hard voting



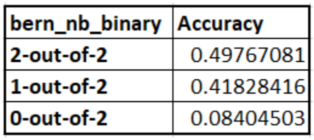
Methods	Min accuracy	Max Accuracy
2-out-of-2	rf_tfidf	nn_binary
1-out-of-2	rf_tf	dt_binary
0-out-of-2	nn_binary	rf_tfidf

Appendix table 146 The methods corresponding to the best and worst results from 4-class classification using hard voting



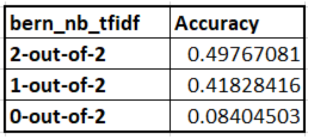
bern_nb.tf	Accuracy
2-out-of-2	0.49767081
1-out-of-2	0.41828416
0-out-of-2	0.08404503

Appendix table 147 Bernoulli NB TF accuracies 4-class classification hard voting



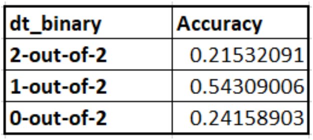
bern_nb_binary	Accuracy
2-out-of-2	0.49767081
1-out-of-2	0.41828416
0-out-of-2	0.08404503

Appendix table 148 Bernoulli NB Binary accuracies 4-class classification hard voting



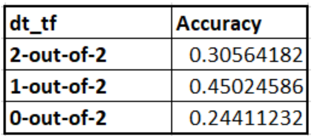
bern_nb_tfidf	Accuracy
2-out-of-2	0.49767081
1-out-of-2	0.41828416
0-out-of-2	0.08404503

Appendix table 149 Bernoulli NB TF-IDF accuracies 4-class classification hard voting



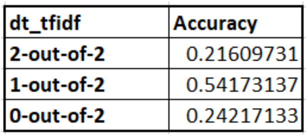
dt_binary	Accuracy
2-out-of-2	0.21532091
1-out-of-2	0.54309006
0-out-of-2	0.24158903

Appendix table 150 DT Binary accuracies 4-class classification hard voting



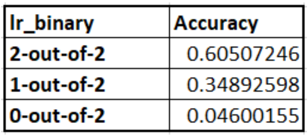
dt_tf	Accuracy
2-out-of-2	0.30564182
1-out-of-2	0.45024586
0-out-of-2	0.24411232

Appendix table 151 DT TF accuracies 4-class classification hard voting



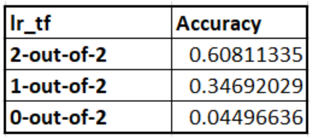
dt_tfidf	Accuracy
2-out-of-2	0.21609731
1-out-of-2	0.54173137
0-out-of-2	0.24217133

Appendix table 152 DT TF-IDF accuracies 4-class classification hard voting



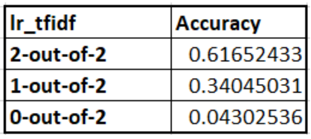
lr_binary	Accuracy
2-out-of-2	0.60507246
1-out-of-2	0.34892598
0-out-of-2	0.04600155

Appendix table 153 LR Binary accuracies 4-class classification hard voting



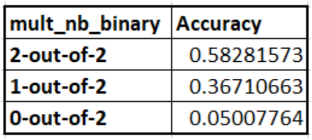
lr_tf	Accuracy
2-out-of-2	0.60811335
1-out-of-2	0.34692029
0-out-of-2	0.04496636

Appendix table 154 LR TF accuracies 4-class classification hard voting



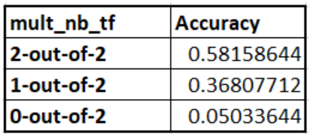
lr_tfidf	Accuracy
2-out-of-2	0.61652433
1-out-of-2	0.34045031
0-out-of-2	0.04302536

Appendix table 155 LR TF-IDF accuracies 4-class classification hard voting



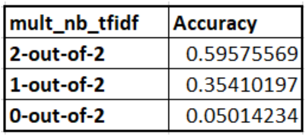
mult_nb_binary	Accuracy
2-out-of-2	0.58281573
1-out-of-2	0.36710663
0-out-of-2	0.05007764

Appendix table 156 Multinomial NB Binary accuracies 4-class classification hard voting



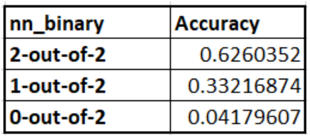
mult_nb_tf	Accuracy
2-out-of-2	0.58158644
1-out-of-2	0.36807712
0-out-of-2	0.05033644

Appendix table 157 Multinomial NB TF accuracies 4-class classification hard voting



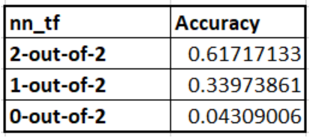
mult_nb_tfidf	Accuracy
2-out-of-2	0.59575569
1-out-of-2	0.35410197
0-out-of-2	0.05014234

Appendix table 158 Multinomial NB TF-IDF accuracies 4-class classification hard voting



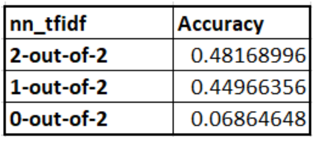
nn_binary	Accuracy
2-out-of-2	0.6260352
1-out-of-2	0.33216874
0-out-of-2	0.04179607

Appendix table 159 NN Binary accuracies 4-class classification hard voting



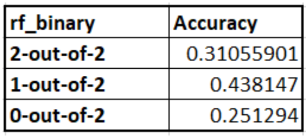
nn_tf	Accuracy
2-out-of-2	0.61717133
1-out-of-2	0.33973861
0-out-of-2	0.04309006

Appendix table 160 NN TF accuracies 4-class classification hard voting



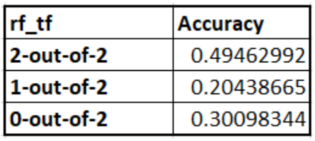
nn_tfidf	Accuracy
2-out-of-2	0.48168996
1-out-of-2	0.44966356
0-out-of-2	0.06864648

Appendix table 161 NN TF-IDF accuracies 4-class classification hard voting



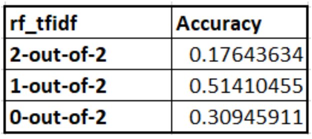
rf_binary	Accuracy
2-out-of-2	0.31055901
1-out-of-2	0.438147
0-out-of-2	0.251294

Appendix table 162 RF Binary accuracies 4-class classification hard voting



rf_tf	Accuracy
2-out-of-2	0.49462992
1-out-of-2	0.20438665
0-out-of-2	0.30098344

Appendix table 163 RF TF accuracies 4-class classification hard voting



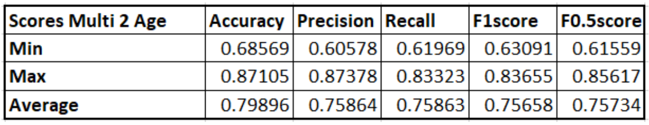
rf_tfidf	Accuracy
2-out-of-2	0.17643634
1-out-of-2	0.51410455
0-out-of-2	0.30945911

Appendix table 164 RF TF-IDF accuracies 4-class classification hard voting

A.3. Results of the combined 4-class data classifiers

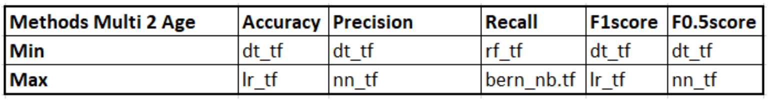
A.3.1. Age classifier

A.3.1.1. Method I (soft vote)



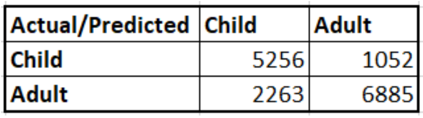
Scores Multi 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.68569	0.60578	0.61969	0.63091	0.61559
Max	0.87105	0.87378	0.83323	0.83655	0.85617
Average	0.79896	0.75864	0.75863	0.75658	0.75734

Appendix table 165 The best and worst results from 4-class combined age classification using soft voting



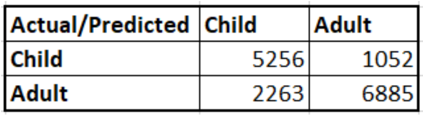
Methods Multi 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	dt_tf	dt_tf	rf_tf	dt_tf	dt_tf
Max	lr_tf	nn_tf	bern_nb.tf	lr_tf	nn_tf

Appendix table 166 The methods corresponding to the best and worst results from 4-class combined age classification using soft voting



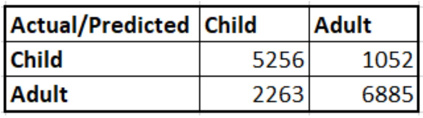
Actual/Predicted	Child	Adult
Child	5256	1052
Adult	2263	6885

Appendix table 167 Bernoulli NB TF on 4-class combined age data soft vote



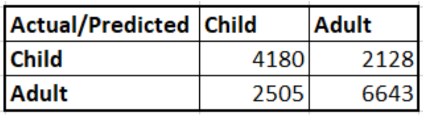
Actual/Predicted	Child	Adult
Child	5256	1052
Adult	2263	6885

Appendix table 168 Bernoulli NB Binary on 4-class combined age data soft vote



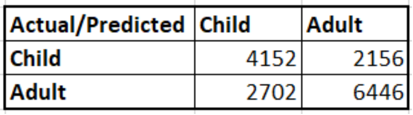
Actual/Predicted	Child	Adult
Child	5256	1052
Adult	2263	6885

Appendix table 169 Bernoulli NB TF-IDF on 4-class combined age data soft vote



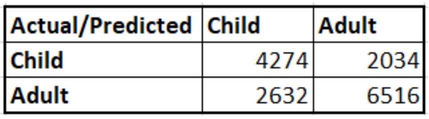
Actual/Predicted	Child	Adult
Child	4180	2128
Adult	2505	6643

Appendix table 170 DT Binary on 4-class combined age data soft vote



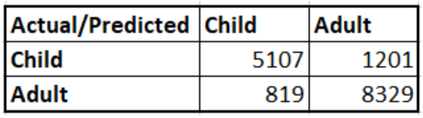
Actual/Predicted	Child	Adult
Child	4152	2156
Adult	2702	6446

Appendix table 171 DT TF on 4-class combined age data soft vote



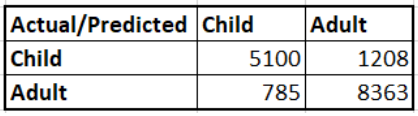
Actual/Predicted	Child	Adult
Child	4274	2034
Adult	2632	6516

Appendix table 172 DT TF-IDF on 4-class combined age data soft vote



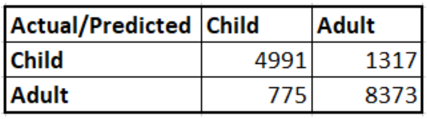
Actual/Predicted	Child	Adult
Child	5107	1201
Adult	819	8329

Appendix table 173 LR Binary on 4-class combined age data soft vote



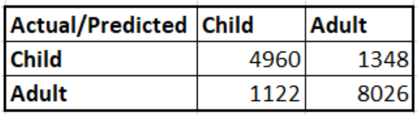
Actual/Predicted	Child	Adult
Child	5100	1208
Adult	785	8363

Appendix table 174 LR TF on 4-class combined age data soft vote



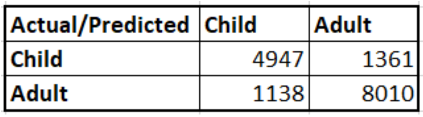
Actual/Predicted	Child	Adult
Child	4991	1317
Adult	775	8373

Appendix table 175 LR TF-IDF on 4-class combined age data soft vote



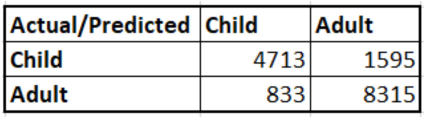
Actual/Predicted	Child	Adult
Child	4960	1348
Adult	1122	8026

Appendix table 176 Multinomial NB Binary on 4-class combined age data soft vote



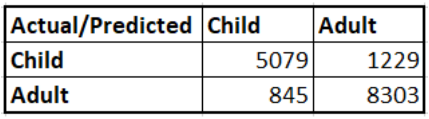
Actual/Predicted	Child	Adult
Child	4947	1361
Adult	1138	8010

Appendix table 177 Multinomial NB TF on 4-class combined age data soft vote



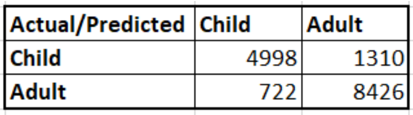
Actual/Predicted	Child	Adult
Child	4713	1595
Adult	833	8315

Appendix table 178 Multinomial NB TF-IDF on 4-class combined age data soft vote



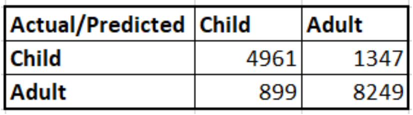
Actual/Predicted	Child	Adult
Child	5079	1229
Adult	845	8303

Appendix table 179 NN Binary on 4-class combined age data soft vote



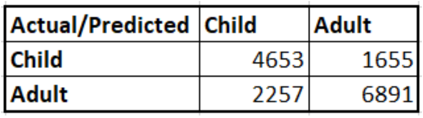
Actual/Predicted	Child	Adult
Child	4998	1310
Adult	722	8426

Appendix table 180 NN TF on 4-class combined age data soft vote



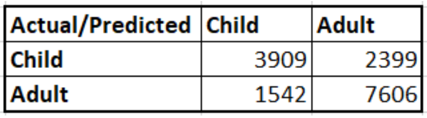
Actual/Predicted	Child	Adult
Child	4961	1347
Adult	899	8249

Appendix table 181 NN TF-IDF on 4-class combined age data soft vote



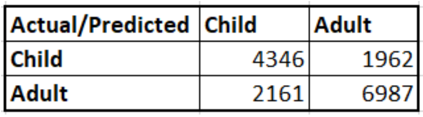
Actual/Predicted	Child	Adult
Child	4653	1655
Adult	2257	6891

Appendix table 182 RF Binary on 4-class combined age data soft vote



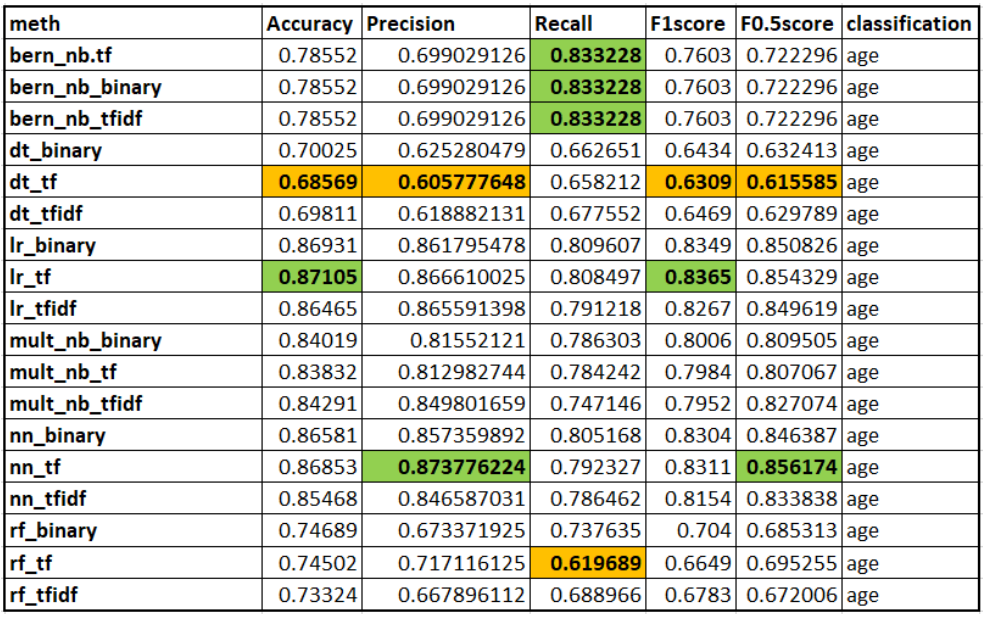
Actual/Predicted	Child	Adult
Child	3909	2399
Adult	1542	7606

Appendix table 183 RF TF on 4-class combined age data soft vote



Actual/Predicted	Child	Adult
Child	4346	1962
Adult	2161	6987

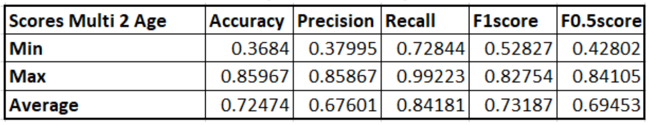
Appendix table 184 RF TF-IDF on 4-class combined age data soft vote



meth	Accuracy	Precision	Recall	F1score	F0.5score	classification
bern_nb.tf	0.78552	0.699029126	0.833228	0.7603	0.722296	age
bern_nb_binary	0.78552	0.699029126	0.833228	0.7603	0.722296	age
bern_nb_tfidf	0.78552	0.699029126	0.833228	0.7603	0.722296	age
dt_binary	0.70025	0.625280479	0.662651	0.6434	0.632413	age
dt_tf	0.68569	0.605777648	0.658212	0.6309	0.615585	age
dt_tfidf	0.69811	0.618882131	0.677552	0.6469	0.629789	age
lr_binary	0.86931	0.861795478	0.809607	0.8349	0.850826	age
lr_tf	0.87105	0.866610025	0.808497	0.8365	0.854329	age
lr_tfidf	0.86465	0.865591398	0.791218	0.8267	0.849619	age
mult_nb_binary	0.84019	0.81552121	0.786303	0.8006	0.809505	age
mult_nb_tf	0.83832	0.812982744	0.784242	0.7984	0.807067	age
mult_nb_tfidf	0.84291	0.849801659	0.747146	0.7952	0.827074	age
nn_binary	0.86581	0.857359892	0.805168	0.8304	0.846387	age
nn_tf	0.86853	0.873776224	0.792327	0.8311	0.856174	age
nn_tfidf	0.85468	0.846587031	0.786462	0.8154	0.833838	age
rf_binary	0.74689	0.673371925	0.737635	0.704	0.685313	age
rf_tf	0.74502	0.717116125	0.619689	0.6649	0.695255	age
rf_tfidf	0.73324	0.667896112	0.688966	0.6783	0.672006	age

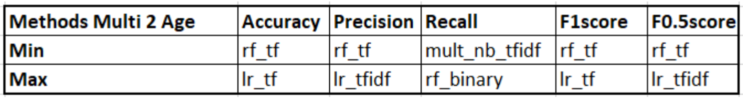
Appendix table 185 All metrics and methods for 4-class combined age classification soft vote

A.3.1.2. Method II (hard vote)



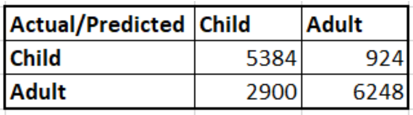
Scores Multi 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.3684	0.37995	0.72844	0.52827	0.42802
Max	0.85967	0.85867	0.99223	0.82754	0.84105
Average	0.72474	0.67601	0.84181	0.73187	0.69453

Appendix table 186 The best and worst results from 4-class combined age classification using hard voting



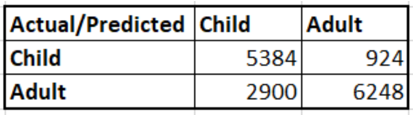
Methods Multi 2 Age	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tf	rf_tf	mult_nb_tfidf	rf_tf	rf_tf
Max	lr_tf	lr_tfidf	rf_binary	lr_tf	lr_tfidf

Appendix table 187 The methods corresponding to the best and worst results from 4-class combined age classification using hard voting



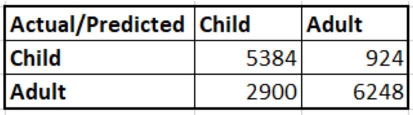
Actual/Predicted	Child	Adult
Child	5384	924
Adult	2900	6248

Appendix table 188 Bernoulli NB TF on 4-class combined age data hard vote



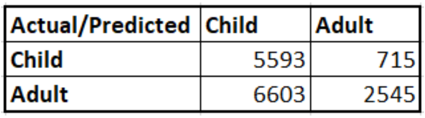
Actual/Predicted	Child	Adult
Child	5384	924
Adult	2900	6248

Appendix table 189 Bernoulli NB Binary on 4-class combined age data hard vote



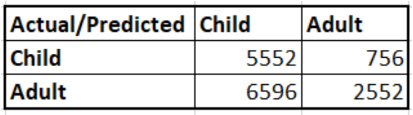
Actual/Predicted	Child	Adult
Child	5384	924
Adult	2900	6248

Appendix table 190 Bernoulli NB TF-IDF on 4-class combined age data hard vote



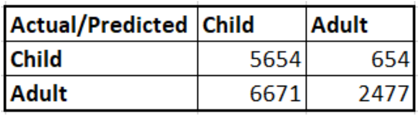
Actual/Predicted	Child	Adult
Child	5593	715
Adult	6603	2545

Appendix table 191 DT Binary on 4-class combined age data hard vote



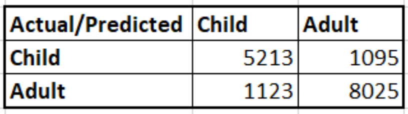
Actual/Predicted	Child	Adult
Child	5552	756
Adult	6596	2552

Appendix table 192 DT TF on 4-class combined age data hard vote



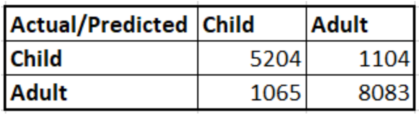
Actual/Predicted	Child	Adult
Child	5654	654
Adult	6671	2477

Appendix table 193 DT TF-IDF on 4-class combined age data hard vote



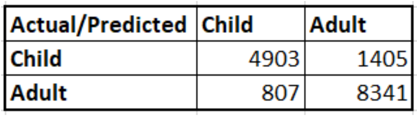
Actual/Predicted	Child	Adult
Child	5213	1095
Adult	1123	8025

Appendix table 194 LR Binary on 4-class combined age data hard vote



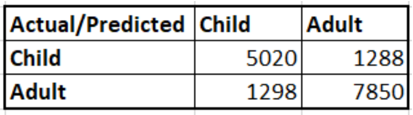
Actual/Predicted	Child	Adult
Child	5204	1104
Adult	1065	8083

Appendix table 195 LR TF on 4-class combined age data hard vote



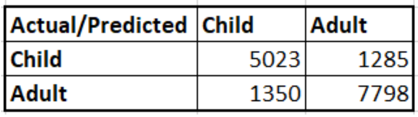
Actual/Predicted	Child	Adult
Child	4903	1405
Adult	807	8341

Appendix table 196 LR TF-IDF on 4-class combined age data hard vote



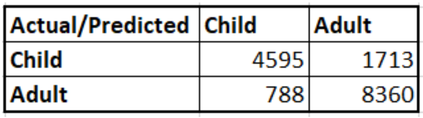
Actual/Predicted	Child	Adult
Child	5020	1288
Adult	1298	7850

Appendix table 197 Multinomial NB Binary on 4-class combined age data hard vote



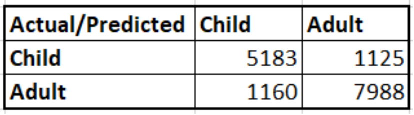
Actual/Predicted	Child	Adult
Child	5023	1285
Adult	1350	7798

Appendix table 198 Multinomial NB TF on 4-class combined age data hard vote



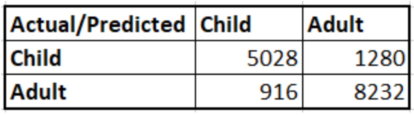
Actual/Predicted	Child	Adult
Child	4595	1713
Adult	788	8360

Appendix table 199 Multinomial NB TF-IDF on 4-class combined age data hard vote



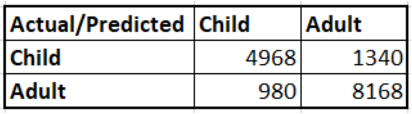
Actual/Predicted	Child	Adult
Child	5183	1125
Adult	1160	7988

Appendix table 200 NN Binary on 4-class combined age data hard vote



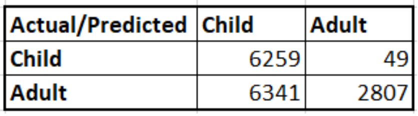
Actual/Predicted	Child	Adult
Child	5028	1280
Adult	916	8232

Appendix table 201 NN TF on 4-class combined age data hard vote



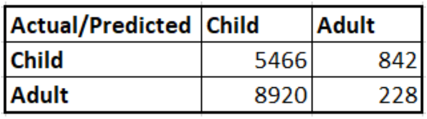
Actual/Predicted	Child	Adult
Child	4968	1340
Adult	980	8168

Appendix table 202 NN TF-IDF on 4-class combined age data hard vote



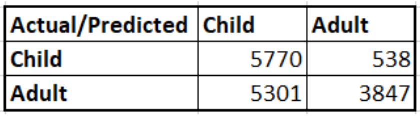
Actual/Predicted	Child	Adult
Child	6259	49
Adult	6341	2807

Appendix table 203 RF Binary on 4-class combined age data hard vote



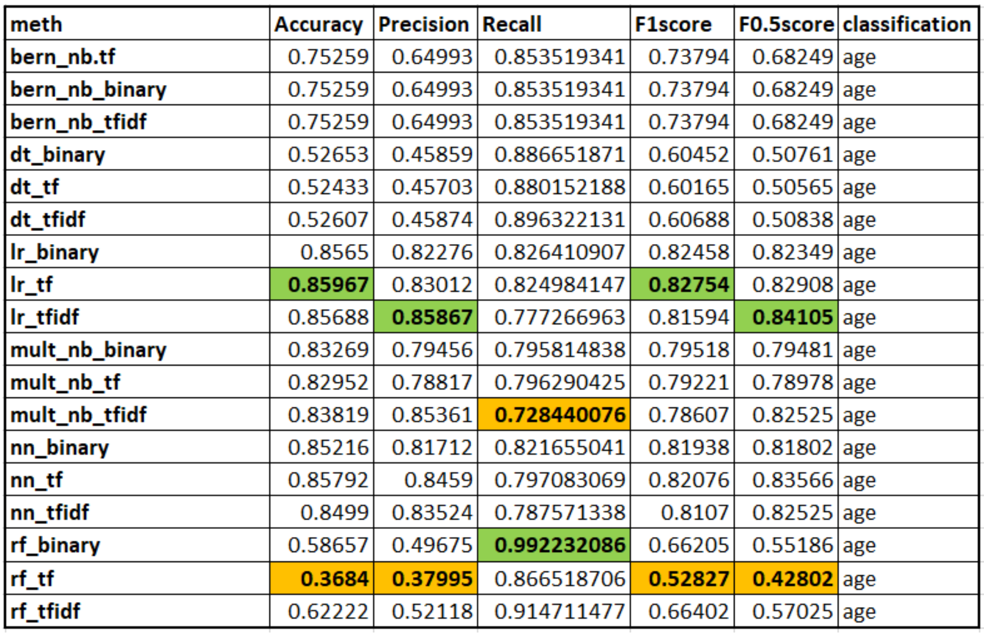
Actual/Predicted	Child	Adult
Child	5466	842
Adult	8920	228

Appendix table 204 RF TF on 4-class combined age data hard vote



Actual/Predicted	Child	Adult
Child	5770	538
Adult	5301	3847

Appendix table 205 RF TF-IDF on 4-class combined age data hard vote

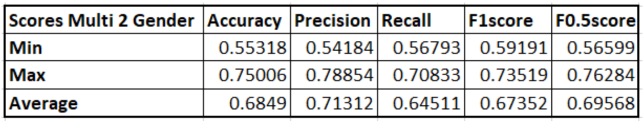


meth	Accuracy	Precision	Recall	F1score	F0.5score	classification
bern_nb.tf	0.75259	0.64993	0.853519341	0.73794	0.68249	age
bern_nb_binary	0.75259	0.64993	0.853519341	0.73794	0.68249	age
bern_nb_tfidf	0.75259	0.64993	0.853519341	0.73794	0.68249	age
dt_binary	0.52653	0.45859	0.886651871	0.60452	0.50761	age
dt_tf	0.52433	0.45703	0.880152188	0.60165	0.50565	age
dt_tfidf	0.52607	0.45874	0.896322131	0.60688	0.50838	age
lr_binary	0.8565	0.82276	0.826410907	0.82458	0.82349	age
lr_tf	**0.85967**	0.83012	0.824984147	**0.82754**	0.82908	age
lr_tfidf	0.85688	**0.85867**	0.777266963	0.81594	**0.84105**	age
mult_nb_binary	0.83269	0.79456	0.795814838	0.79518	0.79481	age
mult_nb_tf	0.82952	0.78817	0.796290425	0.79221	0.78978	age
mult_nb_tfidf	0.83819	0.85361	**0.728440076**	0.78607	0.82525	age
nn_binary	0.85216	0.81712	0.821655041	0.81938	0.81802	age
nn_tf	0.85792	0.8459	0.797083069	0.82076	0.83566	age
nn_tfidf	0.8499	0.83524	0.787571338	0.8107	0.82525	age
rf_binary	0.58657	0.49675	**0.992232086**	0.66205	0.55186	age
rf_tf	**0.3684**	**0.37995**	0.866518706	**0.52827**	**0.42802**	age
rf_tfidf	0.62222	0.52118	0.914711477	0.66402	0.57025	age

Appendix table 206 All metrics and methods for 4-class combined age classification hard vote

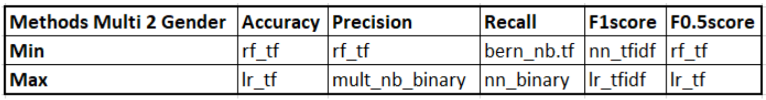
A.3.2. Gender classifier

A.3.2.1. Method I (soft vote)



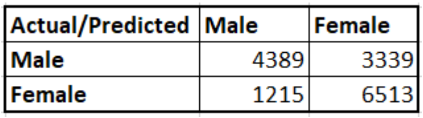
Scores Multi 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.55318	0.54184	0.56793	0.59191	0.56599
Max	0.75006	0.78854	0.70833	0.73519	0.76284
Average	0.6849	0.71312	0.64511	0.67352	0.69568

Appendix table 207 The best and worst results from 4-class combined gender classification using soft voting



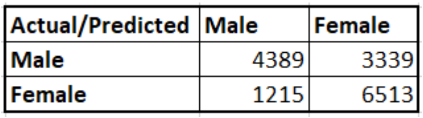
Methods Multi 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tf	rf_tf	bern_nb.tf	nn_tfidf	rf_tf
Max	lr_tf	mult_nb_binary	nn_binary	lr_tfidf	lr_tf

Appendix table 208 The methods corresponding to the best and worst results from 4-class combined gender classification using soft voting



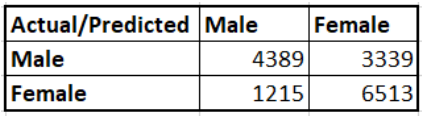
Actual/Predicted	Male	Female
Male	4389	3339
Female	1215	6513

Appendix table 209 Bernoulli NB TF on 4-class combined gender data soft vote



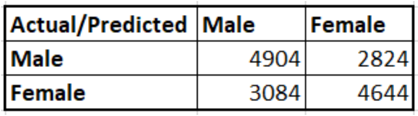
Actual/Predicted	Male	Female
Male	4389	3339
Female	1215	6513

Appendix table 210 Bernoulli NB Binary on 4-class combined gender data soft vote



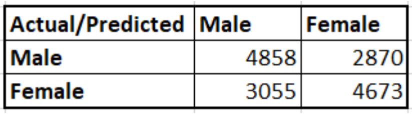
Actual/Predicted	Male	Female
Male	4389	3339
Female	1215	6513

Appendix table 211 Bernoulli NB TF-IDF on 4-class combined gender data soft vote



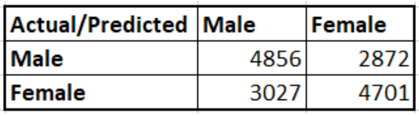
Actual/Predicted	Male	Female
Male	4904	2824
Female	3084	4644

Appendix table 212 DT Binary on 4-class combined gender data soft vote



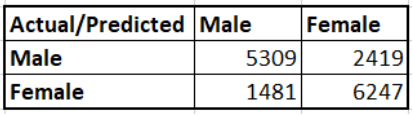
Actual/Predicted	Male	Female
Male	4858	2870
Female	3055	4673

Appendix table 213 DT TF on 4-class combined gender data soft vote



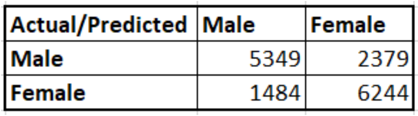
Actual/Predicted	Male	Female
Male	4856	2872
Female	3027	4701

Appendix table 214 DT TF-IDF on 4-class combined gender data soft vote



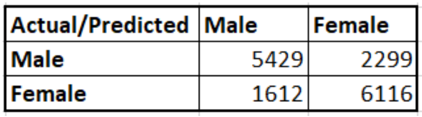
Actual/Predicted	Male	Female
Male	5309	2419
Female	1481	6247

Appendix table 215 LR Binary on 4-class combined gender data soft vote



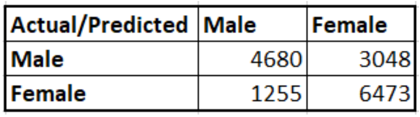
Actual/Predicted	Male	Female
Male	5349	2379
Female	1484	6244

Appendix table 216 LR TF on 4-class combined gender data soft vote



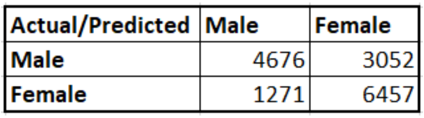
Actual/Predicted	Male	Female
Male	5429	2299
Female	1612	6116

Appendix table 217 LR TF-IDF on 4-class combined gender data soft vote



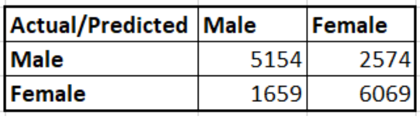
Actual/Predicted	Male	Female
Male	4680	3048
Female	1255	6473

Appendix table 218 Multinomial NB Binary on 4-class combined gender data soft vote



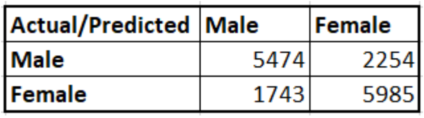
Actual/Predicted	Male	Female
Male	4676	3052
Female	1271	6457

Appendix table 219 Multinomial NB TF on 4-class combined gender data soft vote



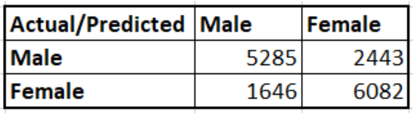
Actual/Predicted	Male	Female
Male	5154	2574
Female	1659	6069

Appendix table 220 Multinomial NB TF-IDF on 4-class combined gender data soft vote



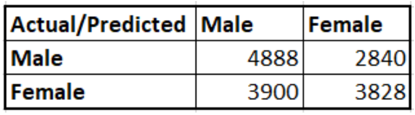
Actual/Predicted	Male	Female
Male	5474	2254
Female	1743	5985

Appendix table 221 NN Binary on 4-class combined gender data soft vote



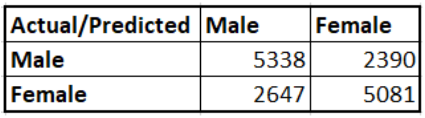
Actual/Predicted	Male	Female
Male	5285	2443
Female	1646	6082

Appendix table 222 NN TF on 4-class combined gender data soft vote



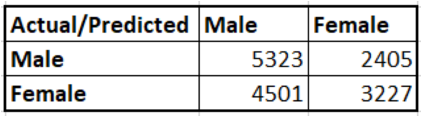
Actual/Predicted	Male	Female
Male	4888	2840
Female	3900	3828

Appendix table 223 NN TF-IDF on 4-class combined gender data soft vote



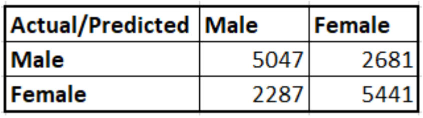
Actual/Predicted	Male	Female
Male	5338	2390
Female	2647	5081

Appendix table 224 RF Binary on 4-class combined gender data soft vote



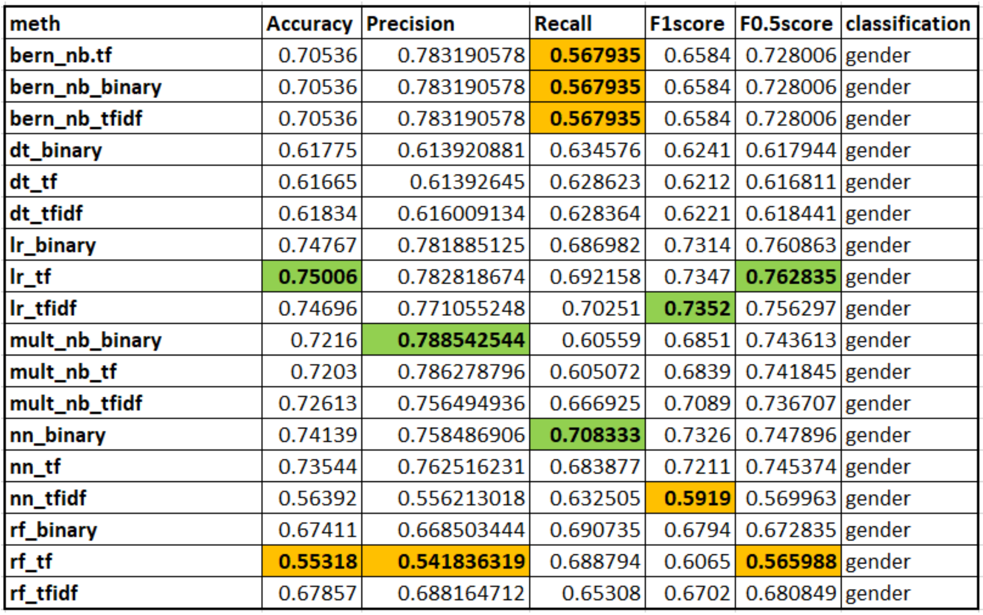
Actual/Predicted	Male	Female
Male	5323	2405
Female	4501	3227

Appendix table 225 RF TF on 4-class combined gender data soft vote



Actual/Predicted	Male	Female
Male	5047	2681
Female	2287	5441

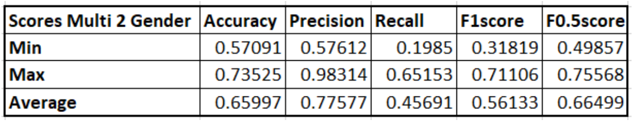
Appendix table 226 RF TF-IDF on 4-class combined gender data soft vote



meth	Accuracy	Precision	Recall	F1score	F0.5score	classification
bern_nb.tf	0.70536	0.783190578	**0.567935**	0.6584	0.728006	gender
bern_nb_binary	0.70536	0.783190578	**0.567935**	0.6584	0.728006	gender
bern_nb_tfidf	0.70536	0.783190578	**0.567935**	0.6584	0.728006	gender
dt_binary	0.61775	0.613920881	0.634576	0.6241	0.617944	gender
dt_tf	0.61665	0.61392645	0.628623	0.6212	0.616811	gender
dt_tfidf	0.61834	0.616009134	0.628364	0.6221	0.618441	gender
lr_binary	0.74767	0.781885125	0.686982	0.7314	0.760863	gender
lr_tf	**0.75006**	0.782818674	0.692158	0.7347	**0.762835**	gender
lr_tfidf	0.74696	0.771055248	0.70251	**0.7352**	0.756297	gender
mult_nb_binary	0.7216	**0.788542544**	0.60559	0.6851	0.743613	gender
mult_nb_tf	0.7203	0.786278796	0.605072	0.6839	0.741845	gender
mult_nb_tfidf	0.72613	0.756494936	0.666925	0.7089	0.736707	gender
nn_binary	0.74139	0.758486906	**0.708333**	0.7326	0.747896	gender
nn_tf	0.73544	0.762516231	0.683877	0.7211	0.745374	gender
nn_tfidf	0.56392	0.556213018	0.632505	**0.5919**	0.569963	gender
rf_binary	0.67411	0.668503444	0.690735	0.6794	0.672835	gender
rf_tf	**0.55318**	**0.541836319**	0.688794	0.6065	**0.565988**	gender
rf_tfidf	0.67857	0.688164712	0.65308	0.6702	0.680849	gender

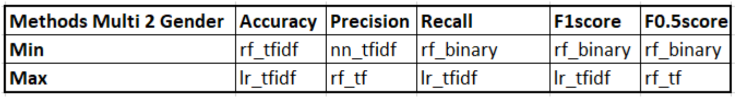
Appendix table 227 All metrics and methods for 4-class combined gender classification soft vote

A.3.2.2. Method II (hard vote)



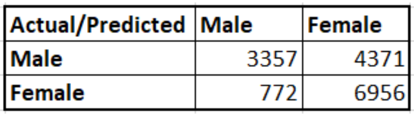
Scores Multi 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	0.57091	0.57612	0.1985	0.31819	0.49857
Max	0.73525	0.98314	0.65153	0.71106	0.75568
Average	0.65997	0.77577	0.45691	0.56133	0.66499

Appendix table 228 The best and worst results from 4-class combined gender classification using hard voting



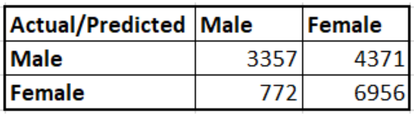
Methods Multi 2 Gender	Accuracy	Precision	Recall	F1score	F0.5score
Min	rf_tfidf	nn_tfidf	rf_binary	rf_binary	rf_binary
Max	lr_tfidf	rf_tf	lr_tfidf	lr_tfidf	rf_tf

Appendix table 229 The methods corresponding to the best and worst results from 4-class combined gender classification using hard voting



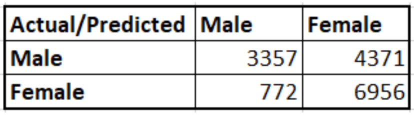
Actual/Predicted	Male	Female
Male	3357	4371
Female	772	6956

Appendix table 230 Bernoulli NB TF on 4-class combined gender data hard vote



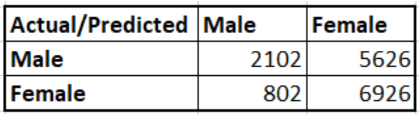
Actual/Predicted	Male	Female
Male	3357	4371
Female	772	6956

Appendix table 231 Bernoulli NB Binary on 4-class combined gender data hard vote



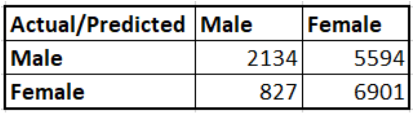
Actual/Predicted	Male	Female
Male	3357	4371
Female	772	6956

Appendix table 232 Bernoulli NB TF-IDF on 4-class combined gender data hard vote



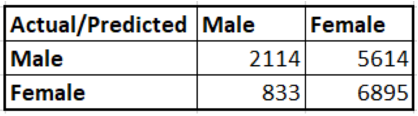
Actual/Predicted	Male	Female
Male	2102	5626
Female	802	6926

Appendix table 233 DT Binary on 4-class combined gender data hard vote



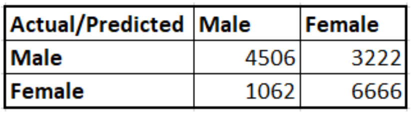
Actual/Predicted	Male	Female
Male	2134	5594
Female	827	6901

Appendix table 234 DT TF on 4-class combined gender data hard vote



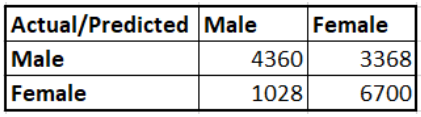
Actual/Predicted	Male	Female
Male	2114	5614
Female	833	6895

Appendix table 235 DT TF-IDF on 4-class combined gender data hard vote



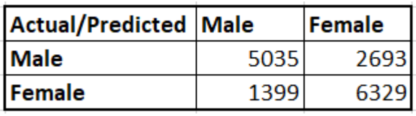
Actual/Predicted	Male	Female
Male	4506	3222
Female	1062	6666

Appendix table 236 LR Binary on 4-class combined gender data hard vote



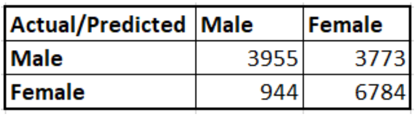
Actual/Predicted	Male	Female
Male	4360	3368
Female	1028	6700

Appendix table 237 LR TF on 4-class combined gender data hard vote



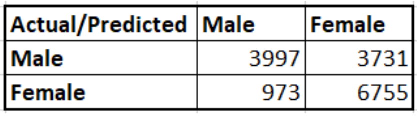
Actual/Predicted	Male	Female
Male	5035	2693
Female	1399	6329

Appendix table 238 LR TF-IDF on 4-class combined gender data hard vote



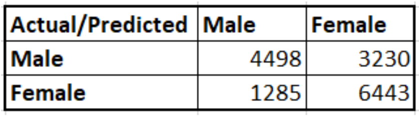
Actual/Predicted	Male	Female
Male	3955	3773
Female	944	6784

Appendix table 239 Multinomial NB Binary on 4-class combined gender data hard vote



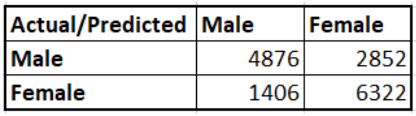
Actual/Predicted	Male	Female
Male	3997	3731
Female	973	6755

Appendix table 240 Multinomial NB TF on 4-class combined gender data hard vote



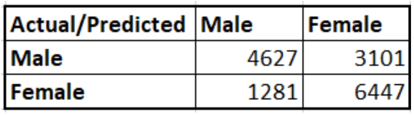
Actual/Predicted	Male	Female
Male	4498	3230
Female	1285	6443

Appendix table 241 Multinomial NB TF-IDF on 4-class combined gender data hard vote



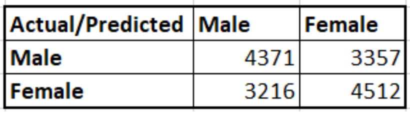
Actual/Predicted	Male	Female
Male	4876	2852
Female	1406	6322

Appendix table 242 NN Binary on 4-class combined gender data hard vote



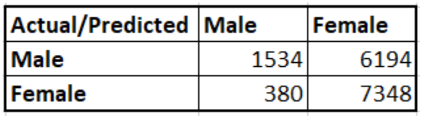
Actual/Predicted	Male	Female
Male	4627	3101
Female	1281	6447

Appendix table 243 NN TF on 4-class combined gender data hard vote



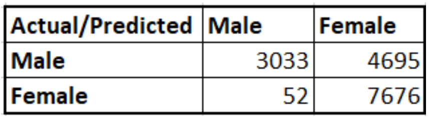
Actual/Predicted	Male	Female
Male	4371	3357
Female	3216	4512

Appendix table 244 NN TF-IDF on 4-class combined gender data hard vote



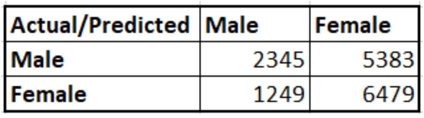
Actual/Predicted	Male	Female
Male	1534	6194
Female	380	7348

Appendix table 245 RF Binary on 4-class combined gender data hard vote



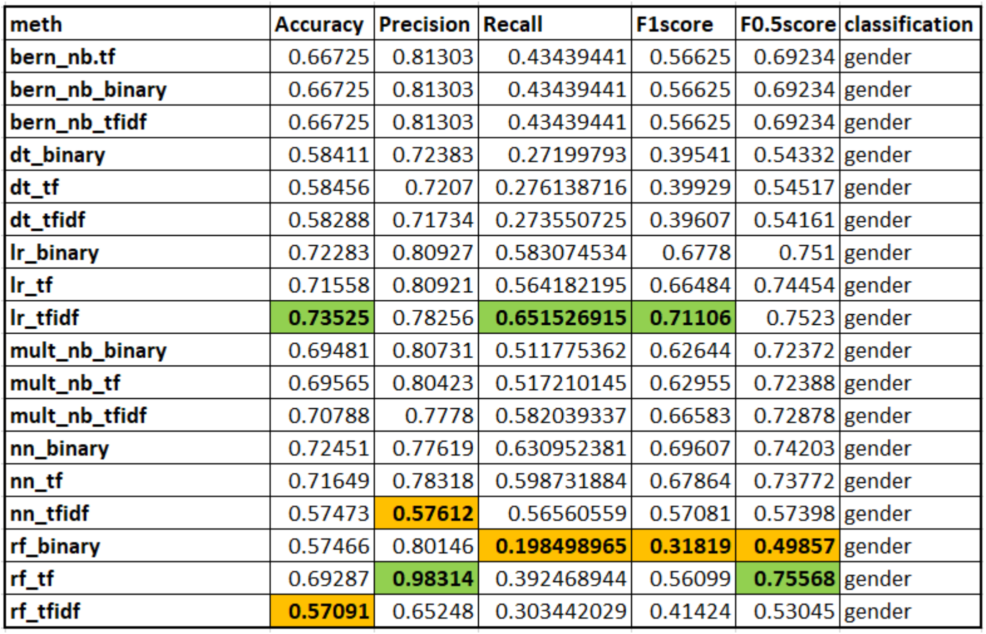
Actual/Predicted	Male	Female
Male	3033	4695
Female	52	7676

Appendix table 246 RF TF on 4-class combined gender data hard vote



Actual/Predicted	**Male**	**Female**
Male	2345	5383
Female	1249	6479

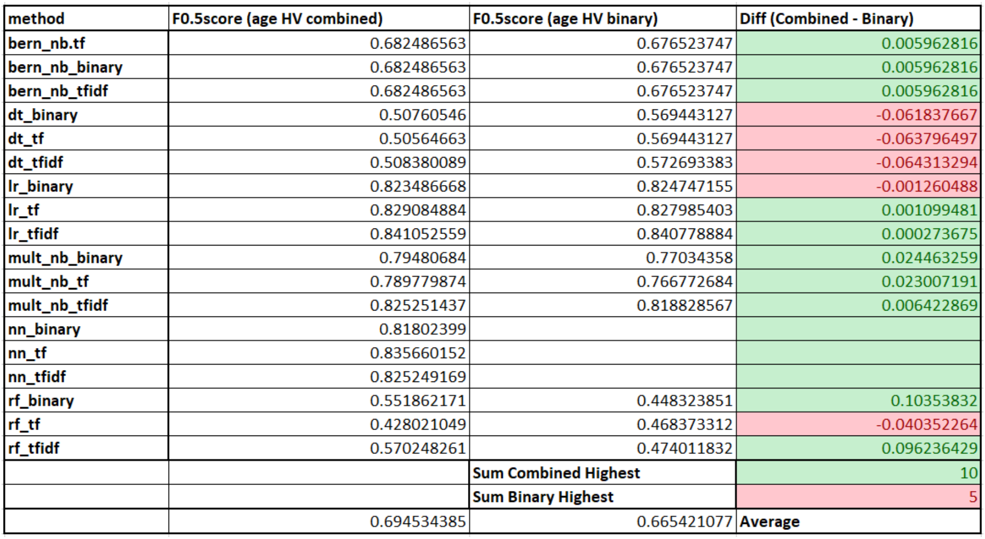
Appendix table 247 RF TF-IDF on 4-class combined gender data hard vote



meth	**Accuracy**	**Precision**	**Recall**	**F1score**	**F0.5score**	**classification**
bern_nb.tf	0.66725	0.81303	0.43439441	0.56625	0.69234	gender
bern_nb_binary	0.66725	0.81303	0.43439441	0.56625	0.69234	gender
bern_nb_tfidf	0.66725	0.81303	0.43439441	0.56625	0.69234	gender
dt_binary	0.58411	0.72383	0.27199793	0.39541	0.54332	gender
dt_tf	0.58456	0.7207	0.276138716	0.39929	0.54517	gender
dt_tfidf	0.58288	0.71734	0.273550725	0.39607	0.54161	gender
lr_binary	0.72283	0.80927	0.583074534	0.6778	0.751	gender
lr_tf	0.71558	0.80921	0.564182195	0.66484	0.74454	gender
lr_tfidf	**0.73525**	0.78256	**0.651526915**	**0.71106**	0.7523	gender
mult_nb_binary	0.69481	0.80731	0.511775362	0.62644	0.72372	gender
mult_nb_tf	0.69565	0.80423	0.517210145	0.62955	0.72388	gender
mult_nb_tfidf	0.70788	0.7778	0.582039337	0.66583	0.72878	gender
nn_binary	0.72451	0.77619	0.630952381	0.69607	0.74203	gender
nn_tf	0.71649	0.78318	0.598731884	0.67864	0.73772	gender
nn_tfidf	0.57473	**0.57612**	0.56560559	0.57081	0.57398	gender
rf_binary	0.57466	0.80146	**0.198498965**	**0.31819**	**0.49857**	gender
rf_tf	0.69287	**0.98314**	0.392468944	0.56099	**0.75568**	gender
rf_tfidf	**0.57091**	0.65248	0.303442029	0.41424	0.53045	gender

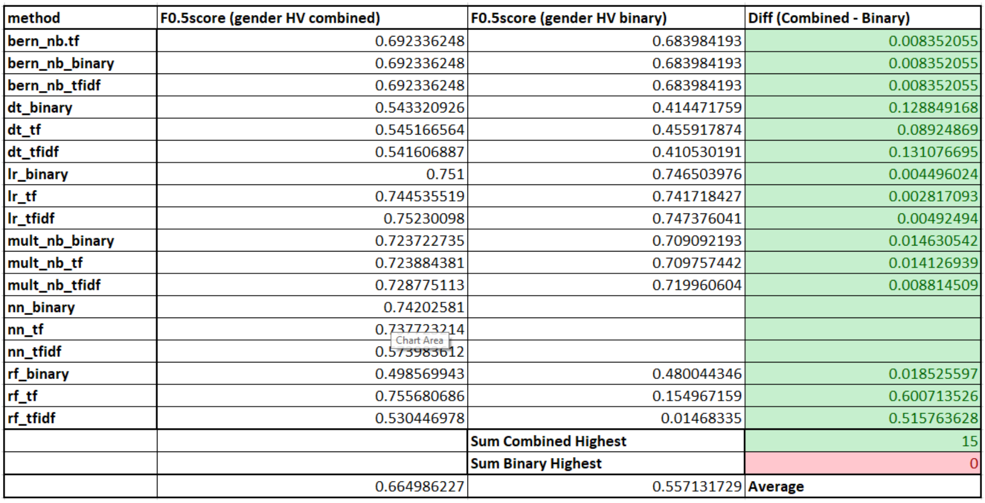
Appendix table 248 All metrics and methods for 4-class combined gender classification hard vote

A.4. Comparison and discussion



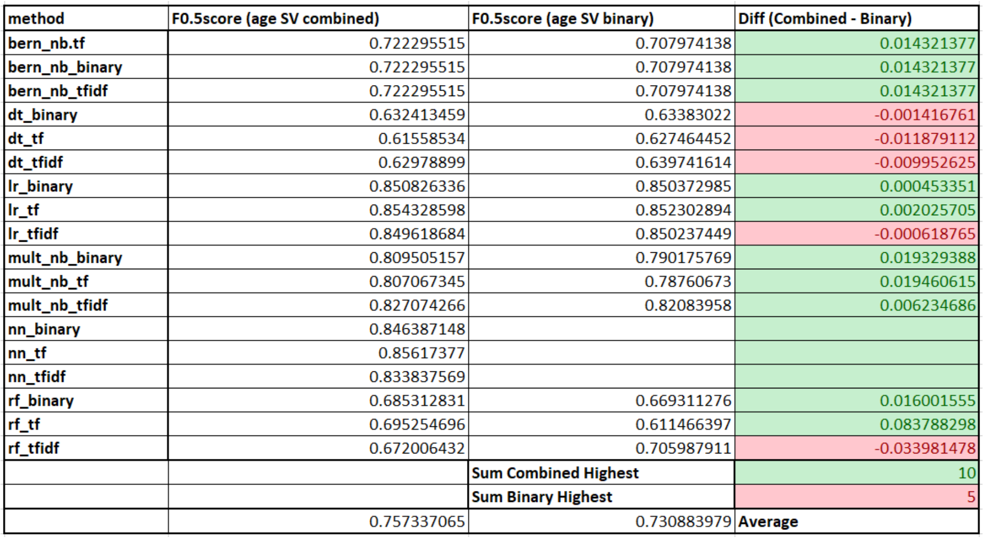
method	F0.5score (age HV combined)	F0.5score (age HV binary)	Diff (Combined - Binary)
bern_nb.tf	0.682486563	0.676523747	0.005962816
bern_nb_binary	0.682486563	0.676523747	0.005962816
bern_nb_tfidf	0.682486563	0.676523747	0.005962816
dt_binary	0.50760546	0.569443127	-0.061837667
dt_tf	0.50564663	0.569443127	-0.063796497
dt_tfidf	0.508380089	0.572693383	-0.064313294
lr_binary	0.823486668	0.824747155	-0.001260488
lr_tf	0.829084884	0.827985403	0.001099481
lr_tfidf	0.841052559	0.840778884	0.000273675
mult_nb_binary	0.79480684	0.77034358	0.024463259
mult_nb_tf	0.789779874	0.766772684	0.023007191
mult_nb_tfidf	0.825251437	0.818828567	0.006422869
nn_binary	0.81802399		
nn_tf	0.835660152		
nn_tfidf	0.825249169		
rf_binary	0.551862171	0.448323851	0.10353832
rf_tf	0.428021049	0.468373312	-0.040352264
rf_tfidf	0.570248261	0.474011832	0.096236429
		Sum Combined Highest	10
		Sum Binary Highest	5
	0.694534385	0.665421077	Average

Appendix table 249 A comparison of the F0.5 scores of all methods within 4-class combined and binary age classification hard vote



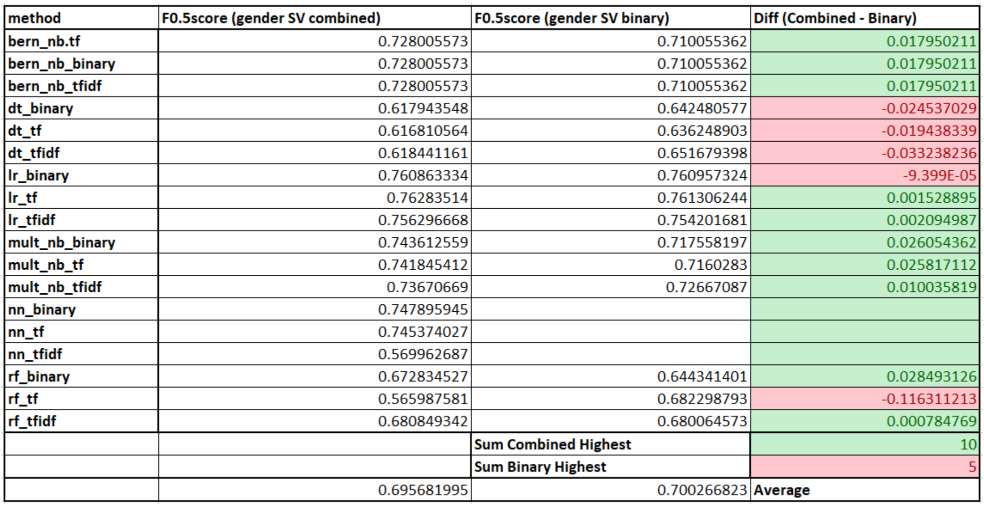
method	F0.5score (gender HV combined)	F0.5score (gender HV binary)	Diff (Combined - Binary)
bern_nb.tf	0.692336248	0.683984193	0.008352055
bern_nb_binary	0.692336248	0.683984193	0.008352055
bern_nb_tfidf	0.692336248	0.683984193	0.008352055
dt_binary	0.543320926	0.414471759	0.128849168
dt_tf	0.545166564	0.455917874	0.08924869
dt_tfidf	0.541606887	0.410530191	0.131076695
lr_binary	0.751	0.746503976	0.004496024
lr_tf	0.744535519	0.741718427	0.002817093
lr_tfidf	0.75230098	0.747376041	0.00492494
mult_nb_binary	0.723722735	0.709092193	0.014630542
mult_nb_tf	0.723884381	0.709757442	0.014126939
mult_nb_tfidf	0.728775113	0.719960604	0.008814509
nn_binary	0.74202581		
nn_tf	0.737723214		
nn_tfidf	0.573983612		
rf_binary	0.498569943	0.480044346	0.018525597
rf_tf	0.755680686	0.154967159	0.600713526
rf_tfidf	0.530446978	0.01468335	0.515763628
		Sum Combined Highest	15
		Sum Binary Highest	0
	0.664986227	0.557131729	Average

Appendix table 250 A comparison of the F0.5 scores of all methods within 4-class combined and binary gender classification hard vote



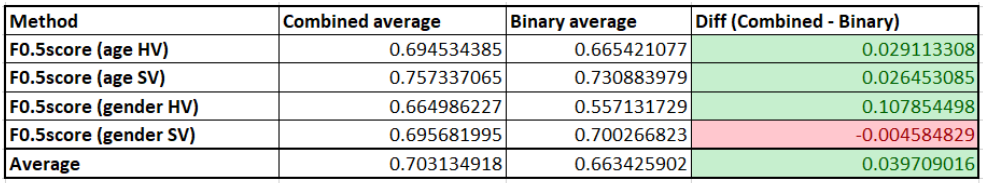
method	F0.5score (age SV combined)	F0.5score (age SV binary)	Diff (Combined - Binary)
bern_nb.tf	0.722295515	0.707974138	0.014321377
bern_nb_binary	0.722295515	0.707974138	0.014321377
bern_nb_tfidf	0.722295515	0.707974138	0.014321377
dt_binary	0.632413459	0.63383022	-0.001416761
dt_tf	0.61558534	0.627464452	-0.011879112
dt_tfidf	0.62978899	0.639741614	-0.009952625
lr_binary	0.850826336	0.850372985	0.000453351
lr_tf	0.854328598	0.852302894	0.002025705
lr_tfidf	0.849618684	0.850237449	-0.000618765
mult_nb_binary	0.809505157	0.790175769	0.019329388
mult_nb_tf	0.807067345	0.78760673	0.019460615
mult_nb_tfidf	0.827074266	0.82083958	0.006234686
nn_binary	0.846387148		
nn_tf	0.85617377		
nn_tfidf	0.833837569		
rf_binary	0.685312831	0.669311276	0.016001555
rf_tf	0.695254696	0.611466397	0.083788298
rf_tfidf	0.672006432	0.705987911	-0.033981478
		Sum Combined Highest	10
		Sum Binary Highest	5
	0.757337065	0.730883979	Average

Appendix table 251 A comparison of the $F_{0.5}$ scores of all methods within 4-class combined and binary age classification soft vote



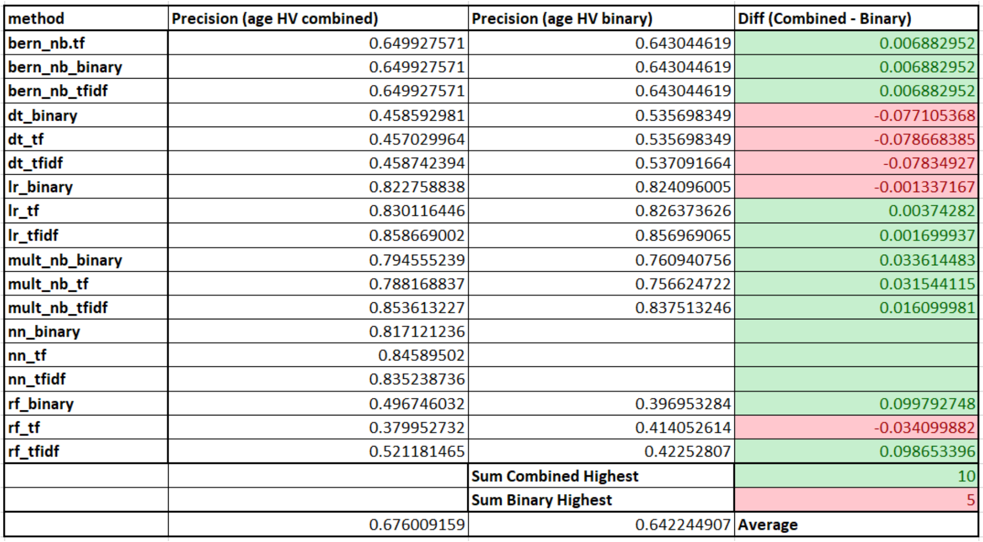
method	F0.5score (gender SV combined)	F0.5score (gender SV binary)	Diff (Combined - Binary)
bern_nb.tf	0.728005573	0.710055362	0.017950211
bern_nb_binary	0.728005573	0.710055362	0.017950211
bern_nb_tfidf	0.728005573	0.710055362	0.017950211
dt_binary	0.617943548	0.642480577	-0.024537029
dt_tf	0.616810564	0.636248903	-0.019438339
dt_tfidf	0.618441161	0.651679398	-0.033238236
lr_binary	0.760863334	0.760957324	-9.399E-05
lr_tf	0.76283514	0.761306244	0.001528895
lr_tfidf	0.756296668	0.754201681	0.002094987
mult_nb_binary	0.743612559	0.717558197	0.026054362
mult_nb_tf	0.741845412	0.7160283	0.025817112
mult_nb_tfidf	0.73670669	0.72667087	0.010035819
nn_binary	0.747895945		
nn_tf	0.745374027		
nn_tfidf	0.569962687		
rf_binary	0.672834527	0.644341401	0.028493126
rf_tf	0.565987581	0.682298793	-0.116311213
rf_tfidf	0.680849342	0.680064573	0.000784769
		Sum Combined Highest	10
		Sum Binary Highest	5
	0.695681995	0.700266823	Average

Appendix table 252 A comparison of the $F_{0.5}$ scores of all methods within 4-class combined and binary gender classification soft vote



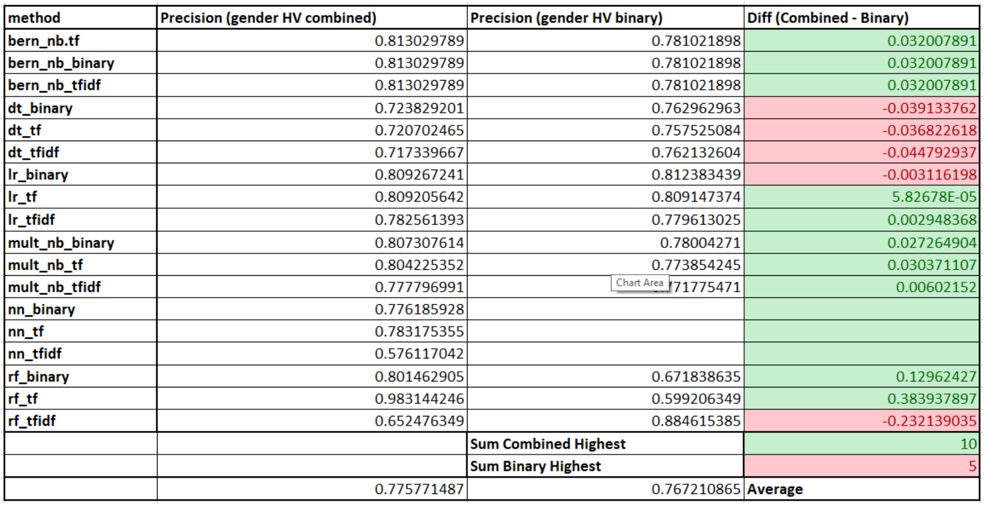
Method	Combined average	Binary average	Diff (Combined - Binary)
F0.5score (age HV)	0.694534385	0.665421077	0.029113308
F0.5score (age SV)	0.757337065	0.730883979	0.026453085
F0.5score (gender HV)	0.664986227	0.557131729	0.107854498
F0.5score (gender SV)	0.695681995	0.700266823	-0.004584829
Average	0.703134918	0.663425902	0.039709016

Appendix table 253 Average $F_{0.5}$ scores compared



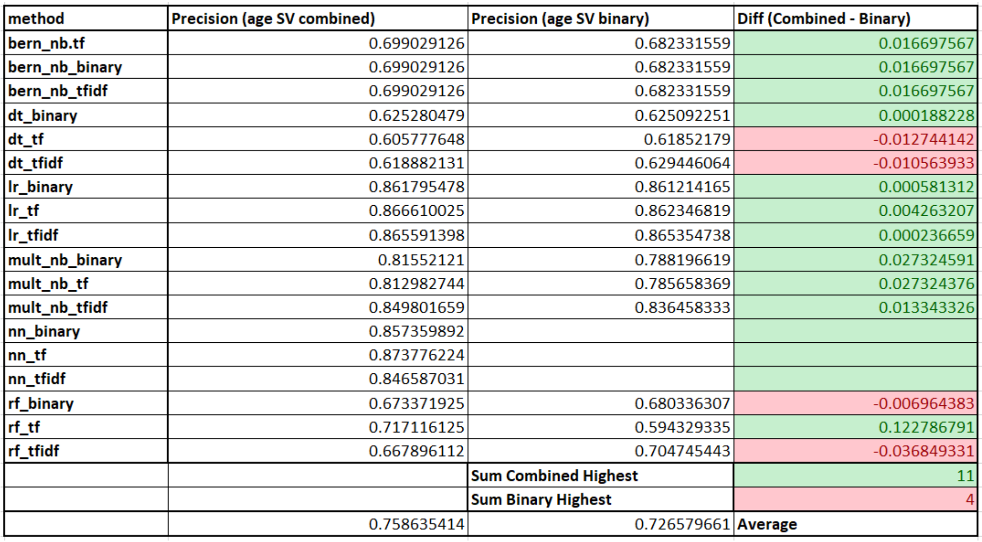
method	Precision (age HV combined)	Precision (age HV binary)	Diff (Combined - Binary)
bern_nb.tf	0.649927571	0.643044619	0.006882952
bern_nb_binary	0.649927571	0.643044619	0.006882952
bern_nb_tfidf	0.649927571	0.643044619	0.006882952
dt_binary	0.458592981	0.535698349	-0.077105368
dt_tf	0.457029964	0.535698349	-0.078668385
dt_tfidf	0.458742394	0.537091664	-0.07834927
lr_binary	0.822758838	0.824096005	-0.001337167
lr_tf	0.830116446	0.826373626	0.00374282
lr_tfidf	0.858669002	0.856969065	0.001699937
mult_nb_binary	0.794555239	0.760940756	0.033614483
mult_nb_tf	0.788168837	0.756624722	0.031544115
mult_nb_tfidf	0.853613227	0.837513246	0.016099981
nn_binary	0.817121236		
nn_tf	0.84589502		
nn_tfidf	0.835238736		
rf_binary	0.496746032	0.396953284	0.099792748
rf_tf	0.379952732	0.414052614	-0.034099882
rf_tfidf	0.521181465	0.42252807	0.098653396
		Sum Combined Highest	10
		Sum Binary Highest	5
	0.676009159	0.642244907	**Average**

Appendix table 254 A comparison of the precision scores of all methods within 4-class combined and binary age classification hard vote



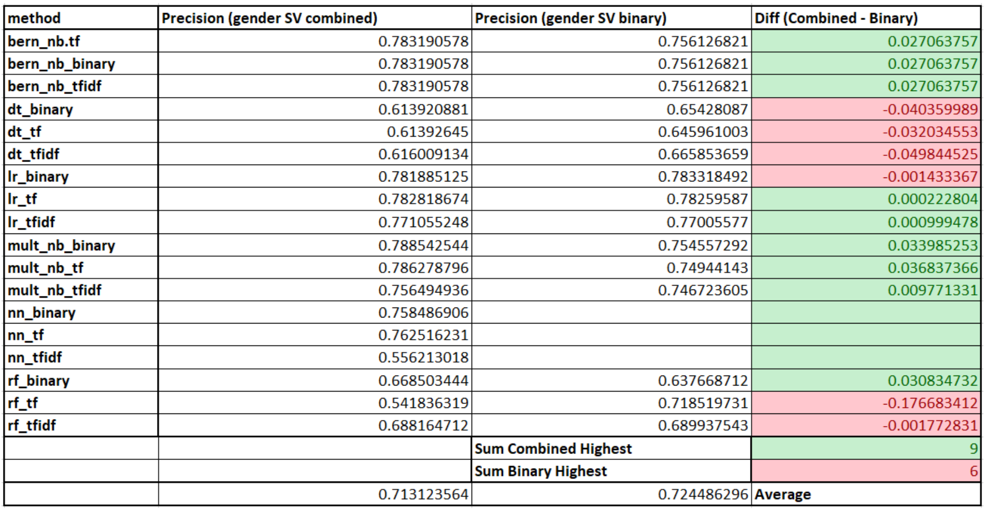
method	Precision (gender HV combined)	Precision (gender HV binary)	Diff (Combined - Binary)
bern_nb.tf	0.813029789	0.781021898	0.032007891
bern_nb_binary	0.813029789	0.781021898	0.032007891
bern_nb_tfidf	0.813029789	0.781021898	0.032007891
dt_binary	0.723829201	0.762962963	-0.039133762
dt_tf	0.720702465	0.757525084	-0.036822618
dt_tfidf	0.717339667	0.762132604	-0.044792937
lr_binary	0.809267241	0.812383439	-0.003116198
lr_tf	0.809205642	0.809147374	5.82678E-05
lr_tfidf	0.782561393	0.779613025	0.002948368
mult_nb_binary	0.807307614	0.78004271	0.027264904
mult_nb_tf	0.804225352	0.773854245	0.030371107
mult_nb_tfidf	0.777796991	Chart Area 71775471	0.00602152
nn_binary	0.776185928		
nn_tf	0.783175355		
nn_tfidf	0.576117042		
rf_binary	0.801462905	0.671838635	0.12962427
rf_tf	0.983144246	0.599206349	0.383937897
rf_tfidf	0.652476349	0.884615385	-0.232139035
		Sum Combined Highest	10
		Sum Binary Highest	5
	0.775771487	0.767210865	**Average**

Appendix table 255 A comparison of the precision scores of all methods within 4-class combined and binary gender classification hard vote



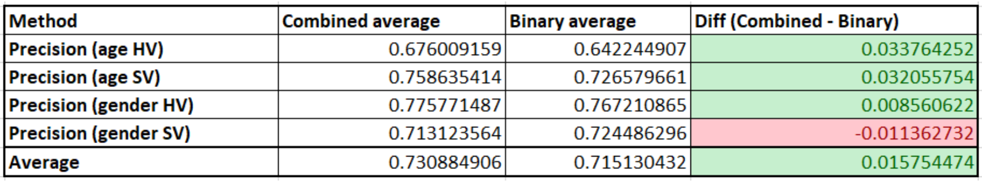
method	Precision (age SV combined)	Precision (age SV binary)	Diff (Combined - Binary)
bern_nb.tf	0.699029126	0.682331559	0.016697567
bern_nb_binary	0.699029126	0.682331559	0.016697567
bern_nb_tfidf	0.699029126	0.682331559	0.016697567
dt_binary	0.625280479	0.625092251	0.000188228
dt_tf	0.605777648	0.61852179	-0.012744142
dt_tfidf	0.618882131	0.629446064	-0.010563933
lr_binary	0.861795478	0.861214165	0.000581312
lr_tf	0.866610025	0.862346819	0.004263207
lr_tfidf	0.865591398	0.865354738	0.000236659
mult_nb_binary	0.81552121	0.788196619	0.027324591
mult_nb_tf	0.812982744	0.785658369	0.027324376
mult_nb_tfidf	0.849801659	0.836458333	0.013343326
nn_binary	0.857359892		
nn_tf	0.873776224		
nn_tfidf	0.846587031		
rf_binary	0.673371925	0.680336307	-0.006964383
rf_tf	0.717116125	0.594329335	0.122786791
rf_tfidf	0.667896112	0.704745443	-0.036849331
		Sum Combined Highest	11
		Sum Binary Highest	4
	0.758635414	0.726579661	Average

Appendix table 256 A comparison of the precision scores of all methods within 4-class combined and binary age classification soft vote



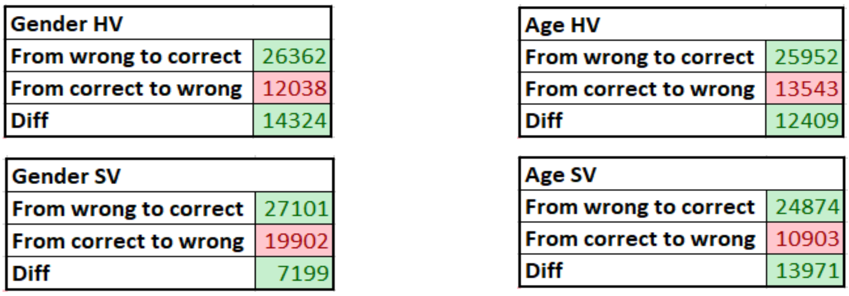
method	Precision (gender SV combined)	Precision (gender SV binary)	Diff (Combined - Binary)
bern_nb.tf	0.783190578	0.756126821	0.027063757
bern_nb_binary	0.783190578	0.756126821	0.027063757
bern_nb_tfidf	0.783190578	0.756126821	0.027063757
dt_binary	0.613920881	0.65428087	-0.040359989
dt_tf	0.61392645	0.645961003	-0.032034553
dt_tfidf	0.616009134	0.665853659	-0.049844525
lr_binary	0.781885125	0.783318492	-0.001433367
lr_tf	0.782818674	0.78259587	0.000222804
lr_tfidf	0.771055248	0.77005577	0.000999478
mult_nb_binary	0.788542544	0.754557292	0.033985253
mult_nb_tf	0.786278796	0.74944143	0.036837366
mult_nb_tfidf	0.756494936	0.746723605	0.009771331
nn_binary	0.758486906		
nn_tf	0.762516231		
nn_tfidf	0.556213018		
rf_binary	0.668503444	0.637668712	0.030834732
rf_tf	0.541836319	0.718519731	-0.176683412
rf_tfidf	0.688164712	0.689937543	-0.001772831
		Sum Combined Highest	9
		Sum Binary Highest	6
	0.713123564	0.724486296	Average

Appendix table 257 A comparison of the precision scores of all methods within 4-class combined and binary gender classification soft vote



Method	Combined average	Binary average	Diff (Combined - Binary)
Precision (age HV)	0.676009159	0.642244907	0.033764252
Precision (age SV)	0.758635414	0.726579661	0.032055754
Precision (gender HV)	0.775771487	0.767210865	0.008560622
Precision (gender SV)	0.713123564	0.724486296	-0.011362732
Average	0.730884906	0.715130432	0.015754474

Appendix table 258 Average precision scores compared



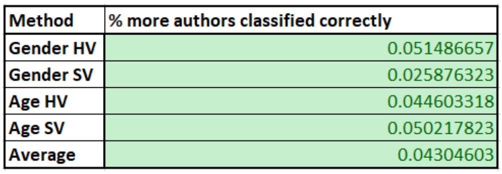
Gender HV	
From wrong to correct	26362
From correct to wrong	12038
Diff	14324

Age HV	
From wrong to correct	25952
From correct to wrong	13543
Diff	12409

Gender SV	
From wrong to correct	27101
From correct to wrong	19902
Diff	7199

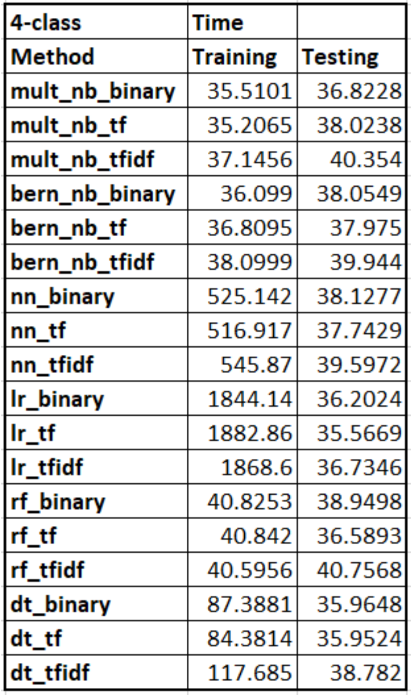
Age SV	
From wrong to correct	24874
From correct to wrong	10903
Diff	13971

Appendix table 259 Differences in classification between 4-class combined and binary classification



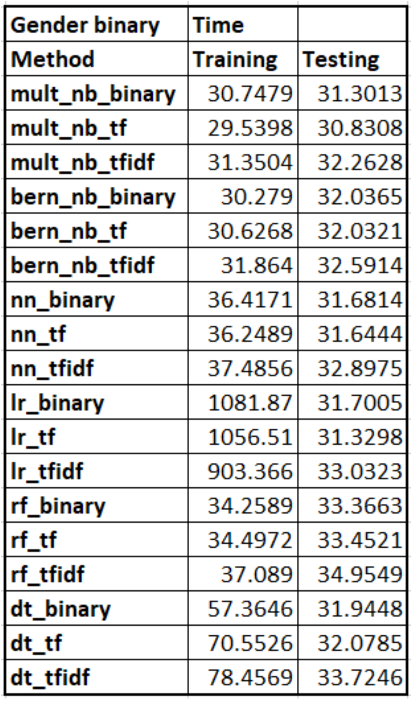
Method	% more authors classified correctly
Gender HV	0.051486657
Gender SV	0.025876323
Age HV	0.044603318
Age SV	0.050217823
Average	0.04304603

Appendix table 260 Difference in correctly classified authors



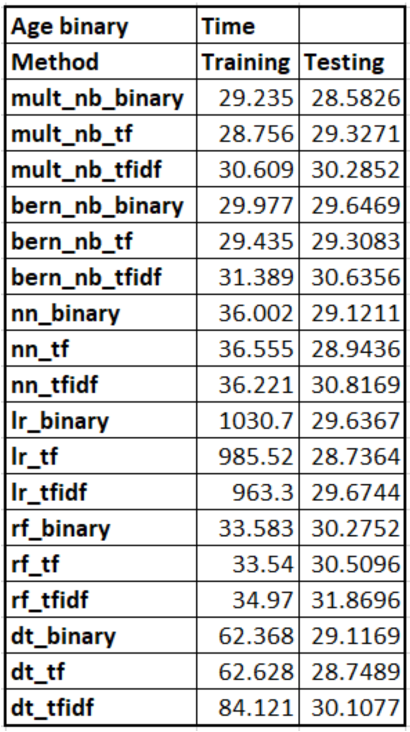
4-class	Time	
Method	Training	Testing
mult_nb_binary	35.5101	36.8228
mult_nb_tf	35.2065	38.0238
mult_nb_tfidf	37.1456	40.354
bern_nb_binary	36.099	38.0549
bern_nb_tf	36.8095	37.975
bern_nb_tfidf	38.0999	39.944
nn_binary	525.142	38.1277
nn_tf	516.917	37.7429
nn_tfidf	545.87	39.5972
lr_binary	1844.14	36.2024
lr_tf	1882.86	35.5669
lr_tfidf	1868.6	36.7346
rf_binary	40.8253	38.9498
rf_tf	40.842	36.5893
rf_tfidf	40.5956	40.7568
dt_binary	87.3881	35.9648
dt_tf	84.3814	35.9524
dt_tfidf	117.685	38.782

Appendix table 261 Computing times for 4-class classification in seconds



Gender binary	Time	
Method	Training	Testing
mult_nb_binary	30.7479	31.3013
mult_nb_tf	29.5398	30.8308
mult_nb_tfidf	31.3504	32.2628
bern_nb_binary	30.279	32.0365
bern_nb_tf	30.6268	32.0321
bern_nb_tfidf	31.864	32.5914
nn_binary	36.4171	31.6814
nn_tf	36.2489	31.6444
nn_tfidf	37.4856	32.8975
lr_binary	1081.87	31.7005
lr_tf	1056.51	31.3298
lr_tfidf	903.366	33.0323
rf_binary	34.2589	33.3663
rf_tf	34.4972	33.4521
rf_tfidf	37.089	34.9549
dt_binary	57.3646	31.9448
dt_tf	70.5526	32.0785
dt_tfidf	78.4569	33.7246

Appendix table 262 Computing times for binary gender classification in seconds



Age binary	Time	
Method	Training	Testing
mult_nb_binary	29.235	28.5826
mult_nb_tf	28.756	29.3271
mult_nb_tfidf	30.609	30.2852
bern_nb_binary	29.977	29.6469
bern_nb_tf	29.435	29.3083
bern_nb_tfidf	31.389	30.6356
nn_binary	36.002	29.1211
nn_tf	36.555	28.9436
nn_tfidf	36.221	30.8169
lr_binary	1030.7	29.6367
lr_tf	985.52	28.7364
lr_tfidf	963.3	29.6744
rf_binary	33.583	30.2752
rf_tf	33.54	30.5096
rf_tfidf	34.97	31.8696
dt_binary	62.368	29.1169
dt_tf	62.628	28.7489
dt_tfidf	84.121	30.1077

Appendix table 263 Computing times for binary age classification in seconds

Emiil Kløvvik

Determining the age and gender of an individual based on text classification

NTNU

Norwegian University of
Science and Technology