CHALMERS
UNIVERSITY OF TECHNOLOGY

Gender Inference on Twitter in Swedish Contexts

Master's thesis in Computer Science: Algorithms, languages and logic

HANNA MATÉRNE

Department of Signals and Systems
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2017

Master's thesis 2017

Gender Inference on Twitter in Swedish Contexts

HANNA MATÉRNE

Department of Signals and Systems
Chalmers University of Technology
Gothenburg, Sweden 2017

Gender Inference on Twitter in Swedish Contexts
HANNA MATÉRNE

Examiner: Andreas Fhager, Department of Signals and Systems

Master's Thesis 2017
Department of Signals and Systems
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2017

Gender Inference on Twitter in Swedish Contexts
HANNA MATÉRNE
Department of Signals and Systems
Chalmers University of Technology

Abstract

This master's thesis investigates methods for inferring the gender of Swedish Twitter users. The dataset, which included account information such as Tweets, personal description, username, etc., was retrieved through Twitters's API. In order to train the models, the accounts were classified manually. Vectors are created to represent the Twitter users. The vectors consist of selected features. The extracted features are meta information such as the description of the user, tweets, username and full name. All the models are based on either Naïve Bayes or Support Vector Machine. A classifier is added to a mixed classifier if it achieves an increased accuracy. The highest accuracy obtained was around 82.35%.

Keywords: Gender inference, Machine learning, Twitter

Contents

List of Figures

List of Tables

1

Introduction

Twitter is a news and social networking service which allows users to read and send messages up to 140 characters. These messages are called *tweets*. With its immense user base of over 300 million active users[1], the service influences all parts of public discourse. People, both in academia and industry, have an interest of being able to obtain demographics of the users. All the tweets of a user contain an overwhelming amount of information about the user's interests, what the user likes etc. Since men and women are often very different, this could be used. Therefore, understanding the user base is highly interesting when targeting specific audience segments, e.g men and women, or for personalization concerns. As an example, radio stations could use it to advertise certain products if they know what kind of gender is listening.
Previous research work focus primarily on classifiers which try to predict the gender of English speaking users [8], but there is also a paper where Ciot et al. do a comparing study of Japanese, Indonesian, Turkish and French users [9], but no Nordic languages. This hints that machine learning methodologies are appropriate for other languages than English as well. Moreover, the positive results from the study hints that further attention within this direction is of interest. In particular investigating Swedish contexts (or any other Nordic language) would be of importance since it has not been covered yet. To conclude, there is a need of a classifier which take into account the Swedish Twitter users.

1.1 Aim and problem formulation

The main aim of the thesis is to design and develop an automatic classification system that can infer users' gender of Swedish Twitter users. This is accomplished using their account information and messaging history.
The overall aim of the thesis can be divided into the following goals:

- Analyse whether tweets, the users' selected username, the name of the user and finally the description is useful to differentiate between genders.

- Constructing a mixed classifier made specific for Swedish Twitter accounts. The resulting accuracy of the mixed classifier is compared at every stage a subclassifier is added, where the purpose is to increase the accuracy of the predictions.

[1] https://about.twitter.com/company

- Analyse the behavioral tweeting patterns of users. For example, the frequencies of re-tweets or the intensity of tweeting sessions might be good differentiators between genders.

1.2 Scope and limitations

This thesis aims to present an approach to classify Swedish Twitter accounts. Classification is limited to *SVM* and *Naïve Bayes.* The work will not include unsupervised learning, and the computational complexity of the algorithms is not a priority.

There are some limitations regarding the dataset. There are no available dataset with pre-labelled accounts in Swedish contexts. Therefore the data collection is performed within the thesis work. Twitter Streaming API is limited to stream rates that allow to stream only small part of the total volume of tweets

1.3 Thesis outline

This chapter, which serves as introduction, will be followed by the following chapters. Chapter 2 includes terminology typical for Twitter. Chapter 3 concerns the machine learning theory on which the thesis work relies. It also gives some theory of common evaluation measures. Chapter 4 describes the methodology used in order to perform the project. In chapter 5, the features and models are listed and explained. Chapter 6 presents the evaluation results. Chapter 7 gives a comparison with previous research and presents interesting extensions for future work. Finally, chapter 8 provides a conclusion of the thesis work.

2
Terminology

The following chapter will describe terms that are characteristic to Twitter.

2.1 Tweet

When a message is posted on Twitter, this is called a tweet. These messages are limited to 140 characters and may contain photos, videos and links.

2.2 Retweet

A retweet can be described as a tweet which is posted by a user that from the beginning is posted by another user. Retweets are often abbreviated RT.

2.3 Protected Tweet

Protected tweets can only be seen by users that have been given a permission by the author.

2.4 Timeline

In a user's timeline, tweets posted by the accounts he or she follows will appear in chronological order.

2.5 Follow

A Twitter user can follow other accounts. When the user follows another account, the tweets by that account will appear in the user's timeline.

2.6 Follower

A user can have followers, i.e accounts that follow that specific user.

2.7 Hashtag

A hashtag '#' is a keyword appointed to a segment of information in order to categorize tweets. In this way it becomes easier to find tweets of specific topics.

2.8 Mention

A mention contains an '@' followed by a person's username anywhere in a tweet.

3

Theory

The purpose of the following chapter is to introduce the theory behind the concepts and methods used in the thesis work. The chapter begins with an introduction to machine learning. Continuing, an overview of two machine learning algorithms is given. Finally, the chapter ends with describing methods for evaluating the performance of the algorithms.

3.1 Machine Learning Theory

This section introduces essential concepts of machine learning. The proposed method of conducting the thesis relies on these ideas. Many issues faced by machine learning can be solved using either classification. These methods are categorizing data, which is what we want to achieve in the end of this thesis, i.e binary answers ("Female" or "Male"). Within machine learning there are an extensive number of algorithms which make it possible to interpret and learn from the data and finally, create a perception or prediction about something. This thesis uses the so called algorithms *Support Vector Machine* and *Naïve Bayes* which will be further described.

Machine learning is a field in computer science which according to Arthur Samuel [10] "gives computers the ability to learn without being explicitly programmed". Machine learning can be divided into several areas; classification, regression and clustering. The goal of classification is to identify to which category or label an object belongs to. When there are two labels only, the term is binary classification. An easy example is, a new observation (an account of Twitter) is classified either as female or male.

Further on, classification can be divided into supervised and unsupervised models. Supervised models are observations and corresponding labels with the aim of recognizing patterns in order to assign labels to new observations. While in unsupervised learning, the models are presented the observations with the aim of recognizing the patterns instead. In the thesis, supervised classification models are used.

3.1.1 Support Vector Machines

Support vector machine, often abbreviated SVM, is a supervised learning algorithm for classification and regression analysis [11].

Given a set of n training examples $x_i \in R^d$, where each training example is marked as either belonging to one or the other out of two categories. The classification is performed by finding the hyper-plane $w^T x - b = 0$ that separates these two categories, where b is the bias of the hyperplane.

There are numerous possible ways of separating hyperplanes and the aim is to trace the hyperplane which maximizes the margin between the both categories. This can be expressed as the following optimization problem:

$$min\frac{1}{2}w^T w \qquad s.t\forall y_i : y_i(w^T x_i + b) \geq 1 \tag{3.1}$$

When it happens that the data is linearly separable, it is possible to choose two parallel hyperplanes which have distance between them as large as possible. Data points close to the optimal hyperplane are called support vectors. When the data is linearly separable, a hard margin entails a perfect classification. This is rather unusual though and to force a hard margin gives an overfitted model. This is avoided by relaxing the constraint into the soft margin version of the problem.

The soft margin modifies the previously stated optimization problem in 3.2 to the primal formulation:

$$\begin{aligned} min\frac{1}{2}w^T w + C\sum_{i=1}^{n}\zeta_i \qquad & s.t \quad \forall y_i : y_i(w * x_i + b) \geq 1 - \zeta_i \\ & \zeta_i = max(0, 1 - y_i(w * x_i + b) \end{aligned} \tag{3.2}$$

Since it is computationally unmanageable for higher dimensions to solve the primal formulation in an optimal way, the use of the dual formulation is advocated. This equation can be seen in 3.3.

$$max\sum_{i}^{n}\alpha_i - \frac{1}{2}\sum_{i}^{n}\sum_{j}^{n}y_i y_j \alpha_i \alpha_j (x_i * x_j) \qquad s.t\sum_{i}^{n}y_i\alpha_i = 0, \forall i : 0 \leq \alpha_i \leq C \tag{3.3}$$

The α-values results in a weight vector.

$$w = \sum_{i}^{n}\alpha_i y_i x_i \tag{3.4}$$

A new observation $x_{observation}$ can be predicted when the Support vector machine is trained.

$$predict(x_{observation}) = sgn(w * \gamma(x_{observation}) + b) = sgn([\sum_{i=1}^{n}\alpha_i y_i k(x_i, x_{observation})] + b) \tag{3.5}$$

3.1.2 Naïve Bayes

Naïve Bayes is a simple classification method which is based on Bayes' rule [12]. The text is utilized as a bag of words. Bag of words means that the order of the words in a text does not matter. For a document d and a class c, we have that:

$$\begin{aligned} P(c|d) = \frac{P(d|c)P(c)}{P(d)} c_{MAP} \\ = argmax P(c|d) \\ = argmax \frac{P(d|c)P(c)}{P(d)} \\ = argmax P(d|c)P(c) \\ = argmax P(x_1, x_2, ..., x_n|c)P(c) \end{aligned} \tag{3.6}$$

3.2 Mixed classifier

A mixed classifier, often called combined classifier, is a classifier which combines several classifiers to obtain a new prediction. The advantage of this technique is that the obtained accuracy by the mixed classifier often is higher than without.

3.3 Performance measures

In binary decision problems, the classifier labels the data either as positive or negative. This can be represented as a confusion matrix which has four different classes:

- True positives (TP): The data is correctly labeled as positives.
- False positives (FP): The negative data incorrectly labeled as positive.
- True negatives (TN): The negative data correctly labeled as negative.
- False negatives (FN): The positive data incorrectly labeled as negative.

A confusion matrix can be seen in 3.1.

	Actual positive	Actual negative
Predicted positive	*TP*	*FN*
Predicted negative	*FP*	*TN*

Table 3.1: Classification outcomes of binary decision problems.

The most used performance measures can be derived from the confusion matrix; accuracy, precision and recall.

Accuracy describes the proportion of classified values that were correct.

$$\text{accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}}$$

Precision describes the positive predictive value. It is a measure of the amount of accurate positives the model claims compared to the number of positives it actually claims.

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}}$$

Recall is the true positive rate. It is a measure of the amount of positives the model claims compared to the actual number of positives there are throughout the data.

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}}$$

A common way to present the relationship between precision and recall is by Receiver Operator Characteristic (ROC). But when the data sets are highly skewed, Precision-Recall curves are better in giving a more detailed picture of the algorithm's performance.

When combining precision and recall, the F1-score is obtained. This is defined as the harmonic mean of precision and recall:

$$F1 = 2 * \frac{\text{PRECISION}*\text{RECALL}}{\text{PRECISION+RECALL}}$$

4
Methodology

This chapter outlines the practical methodology used for retrieving the dataset, perform the data preprocessing steps and implementing the classifiers in the work. The chapter begins with a description how the data needed to accomplish the classifications are retrieved. This is followed by the data preprocessing steps. Finally, the chapter ends with an outline of the practical implementation details.

4.1 Dataset

The following section will describe the dataset of Swedish Twitter users. The purpose of this dataset is to train and test the different models in order to make predictions. Since our classifiers are supervised, it will require that the data is labelled. Unfortunately in our case, when a user registers on Twitter, gender information is not included.

4.1.1 Creating labelled data

The data was gathered from Twitter using its streaming APIs. Since the goal is to infer the gender in Swedish contexts, followers of well-known Swedish Twitter accounts were retrieved. These accounts were *SvD*, *dagensnyheter* and *expressen*. The retrieved dataset was manually reviewed, meaning that the following accounts were taken away:

- Accounts of companies or organizations.
- Accounts without any tweets.
- Accounts without any Swedish tweets.
- Accounts with protected tweets.
- Accounts with only retweets.
- Accounts of users where it was impossible to determine the gender.

The rest of the accounts were classified manually either as female or male. Usually, the name and profile picture give enough information in order to determine the gender. But in some cases the user did not have any of these attributes or, occasionally, attributes were not descriptive enough. If the profile picture and name were not

enough, the description was taken into account. It is easy to classify when the description contains words as ’mamma’ (mother), ’pappa’ (father), ’kvinna’ (female) or ’man’ (male). If there were no such clues or no description of the user, the tweets could give relevant information. If it still was impossible to classify whether the account were a female or a male, the user was eliminated from the dataset. It is of great importance that the users is classified correctly. This since the resulting dataset will constitute as training data.
In Table 4.1 an overview of the resulting dataset is given.

Username	Females	Males	Total
SvD	258	344	602
dagensnyheter	568	576	1144
expressen	459	516	975
TOTAL	1285	1436	2721

Table 4.1: Number of retrieved users

4.2 Data preprocessing

For the sake of using the data as input for the different models, several preprocessing steps need to be performed. These steps are needed in order to analyze the data. When the data is not carefully examined, it can entail misleading results. The preprocessing steps include feature extraction and feature selection. These steps are further described in the following subsections. When the preprocessing steps are completed, a final training set is produced.

4.2.1 Feature extraction

The feature extraction stage converts attributes into data points which constitute input to the classification models. Every data point consists of a number of extracted features integrated to a feature vector. The features extracted are further described in 5.1.

4.2.2 Data representation

When classifying the users, the users need a representation in terms of vectors. In other words, the features are brought together in a vector. These vectors constitute input to the classifiers. The classifiers try to find patterns in the values of the vectors. The patterns are the basis for predicting the gender of the users in the dataset.

4.2.3 Feature selection

Since there are millions of features, the classifiers will not be able to check every feature to find patterns. Thus, a selection of features needs to be performed. This is done by computing the mutual information.

4.3 Algorithms

This section will give a presentation of the different algorithms used in the evaluation process.

4.3.1 Support Vector Machine (SVM)

The first algorithm used is SVM. This algorithm is further described in Section 3.1.1. The reason for using this algorithm is because it has been frequently used in several studies [1, 3, 4, 5] which makes it interesting to try out in this context as well. The kernel used is the radial basis function kernel.

4.3.2 Naïve Bayes

The second algorithm used is Naïve Bayes. This algorithm is further described in Section 3.1.2. In a study by Miller et al. [13], the algorithm shows very promising result, where the accuracy always is between 90-100%. Naïve Bayes is easy to implement and requires only small amounts of training data to be able to estimate the parameters. On the other hand, it needs a larger amount of labelled data to train the model.

4.4 Technical information

This section presents the details of the practical implementation. All experiments in the thesis project is written in the programming language Python. The reason for choosing Python is because of its' popularity in scientific contexts. In order to perform the classification the library *scikit-learn* was used.
In order to handle the data in a easier and faster way, a SQLite database was developed.

5
Features and Models

The following chapter will present the attributes of an account used to predict its user's gender. We will also present how these attributes are transformed to features. The features are input to the models, where the underlying machine learning techniques try to find patterns. Thus, the chapter will also describe the features and models.

5.1 Feature Extraction

The upcoming subsections will explain how the the different attributes are turned into features. A feature represents a measurable property of an instance being observed. The features are used by the machine learning techniques to find patterns in the values of the features.

5.1.1 Username

The username is the only attribute that all Twitter users must have. Additionally, the username is unique. Often the user's first name form a part of the username. Thus, when creating substrings, i.e N-grams, of the username it is beneficial to investigate whether it can be useful when classifying. This since it might occur for several users.

5.1.2 Name

Every user does not provide their full name. When registering, it is required to input at least one character for the name. When extracting features from the name, the name is reconstructed as follows: Firstly, every uppercase letter is converted to a lowercase letter. Secondly, Whitespaces are eliminated. Thirdly, characters that are not a letter or space are taken away. Fourthly, accented letters are converted into the corresponding letter without accent.

5.1.3 Description

Every user can provide a description of themselves. A description is optional, and in the obtained dataset about 65% of the users have provided a description. Often the descriptions contain words which can give guidance in deciding the gender. For example, a user can describe himself as "trebarnsfar" (father of three) or describe an interest that is more popular among women than men or vice versa. In order to find these, several conversions need to be carried out. Additionally, the aim is to combine features that could be combined into one and if needed, differ characters to obtain two features.

- Accented letters are converted into the corresponding letter without accent.
- All hashtags are eliminated. The reason is the same as above.
- Commas are eliminated.
- Exclamation and quotation marks are eliminated. The reason is because 'welcome!' and 'welcome' should not be different features.
- Numbers are eliminated.
- Punctuation before or after a number of letters are eliminated.
- Repeated periods are eliminated.
- Tokens of length one are eliminated.
- Uppercase letters are converted into lowercase letters.
- URLs are eliminated.
- Whitespaces are eliminated.
- Words in singular and plural should count as the same word. Therefore, 'er' at the end of words are eliminated.

5.1.4 Tweets

There are several ways to extract features from the tweets. These are described further in the following section. Every feature instance is counted. Our training set consists of 4542 tweets and the test set consists of 2224 tweets.
The first way of extracting features is based on style attributes. The following features are extracted from the text in the tweets:

- Number of capital letters
- Number of emojis
- Number of emoticons
- Number of exclamation marks
- Number of extended letter sequence
- Number of extended sign sequences
- Number of periods
- Number of question marks
- Number of quotations
- Number of words

The other way of extracting features is based on the actual content such as the words and emoticons. The text in the tweets are rearranged in the same way as described in Section 5.1.3.

5.2 Feature representation

The following section will present how the previously described features, see Section 5.1, are represented as vectors. Section 5.2.1 presents a description about vectors in general. Section 5.2.2 describes how the selection of features is done.

5.2.1 What is a vector?

Vectors are used to represent the accounts in order to classify these. A vector consists of n elements if n features are included, $\vec{x_j} = (x_1, x_2, ..., x_n)$. Feature i is represented by x_i. The value of x_i is equal to zero if the corresponding feature does not happen for that specific account j. Otherwise, x_i corresponds to the percentage where the feature occurs. The final vectors constitute input for the models. The values of the vectors are used by the models. And hopefully, the models find patterns that benefits the model when predicting the gender of new accounts.

5.2.2 Feature selection

Feature selection is the process of selecting a subset of relevant features. A selection of features needs to be performed since the number of features are very large and it becomes hard to productively find patterns. It also results in less training time. The selection is done in a few operations.

Depending on what type of vector we handle, each input is transformed by a transformer. The purpose of this is to simplify the counting of features, meaning for example that some tokens can be combined into the same feature. If a feature is not necessary it is eliminated during this operation.

The next operation is called *tokenization*. The purpose is to divide the text into smaller parts, ending with a whitespace. This operation assembles n-grams as well. The result of the tokenization is the features.

As mentioned earlier in this section, the number of features are very large. Thus, it is of great importance to minimize the number of features, partly because of memory consumption but also to make it simpler to recognize patterns amongst the features. A common way to make the selection is to compute the mutual information 5.1, which measures how much information the precence/abscence of a term contributes to making the correct classification decision. Equation 5.1 [7] tells us that N_{11} represents the number of males which have a specific feature and N_{01} represents the number of males which do not. Correspondingly, N_{10} describes the females which have a specific feature and N_{00} represents the number of females which do not. The total number of males is $N_{11} + N_{01}$ and the total number of females is $N_{10} + N_{00}$.

$$
\begin{aligned}
I(U;C) = {} & \frac{N_{11}}{N} \times log_2 \times \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \times log_2 \times \frac{NN_{01}}{N_{0.}N_{.1}} \\
& + \frac{N_{10}}{N} \times log_2 \times \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \times log_2 \times \frac{NN_{00}}{N_{0.}N_{.0}}
\end{aligned} \tag{5.1}
$$

When the features are selected through mutual information, the creation of vectors is performed.

5.3 Models

In the following section, a presentation of the different models we are evaluating is given. In Section 4.1, the creation of the dataset was described. All instances of the dataset is manually classified. The purpose of using this classified dataset is to learn from these instances and to find patterns. When introducing a completely new dataset, the goal is to classify the instances as female or male correctly. A subset of our achieved dataset will constitute as training set.

5.3.1 Name features model

The name features model uses vectors. If the name of the user is an empty string after transformation, the vectors will not be created. Naïve Bayes was adapted to the features. The technique is further explained in 3.1.2. The number of selected features is 3500 which were selected through Equation 5.1. These consisted of the name features which had the most mutual information.

5.3.2 Username features model

The username features model uses vectors. Since username is required, these vectors are always available. The model utilizes Naïve Bayes, further explained in 3.1.2. Using Equation 5.1, 2500 features were selected. These consisted of the username features which had most mutual information.

5.3.3 Description features model

The description features model creates vectors by summing up the features. If the user does not provide any description, a vector will not be created. Thus, this model will only operate for users providing a description. The model utilizes Support Vector Machine, further explained in 3.1.1. Using Equation 5.1, 1000 features were selected. These consisted of the description features which had most mutual information.

5.3.4 Tweet features model

This model creates vectors by summing up the features. If the user has never posted any tweets, a vector will not be created. Thus, this model will only operate for users that have posted tweets.

The first part of this model, which is based on the style attributes, utilizes Support Vector Machine, further explained in 3.1.1. Using Equation 5.1, 100 features were selected. These consisted of features (based on style attributes) which had most mutual information.

The second part of this model, based on the actual words and emoticons, also utilizes Support Vector Machine. Using Equation 5.1, 1000 features were selected. These consisted of features (based on words) which had most mutual information.

6

Evaluation Results

The following chapter presents and highlights the results after running each classifier. Section 6.1 will present the result of the different classifiers compared to each other. Section 6.2 will present the result of the combined classifier, where the goal is to achieve a higher accuracy.

6.1 Comparison of the different models

The following section presents the result of the different components of the combined classifier. The training set consisted of 1500 accounts, i.e 750 females and 750 males. The number of features are described in the previous chapter, see Chapter 5. The figure below illustrates the accuracy of every classifier. From this figure we can conclude that the classifier with the highest accuracy was the name classifier, here abbreviated NC, with a percentage of 80.12%. It is followed by the username classifier, here abbreviated UC, which obtained an accuracy of 76.19%. The tweet word features classifier, here abbreviated WC, obtained an accuracy of 72.32%. The tweet style features classifier, here abbreviated SC, has the second lowest accuracy, 65.34%. This classifier is closely followed by the description classifier, here abbreviated DC, with a percentage of 64.23% in accuracy.

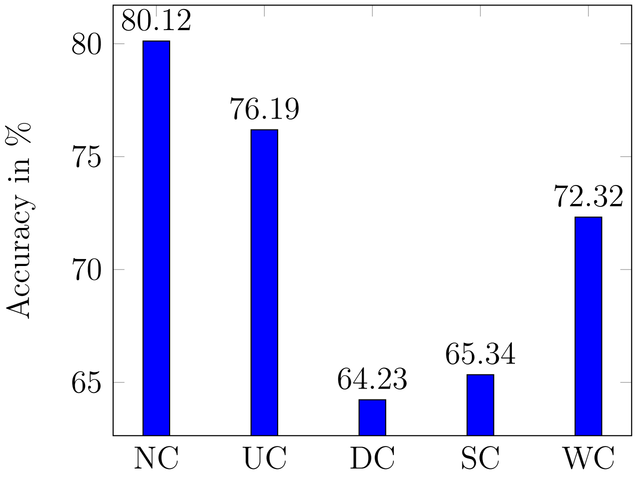

Figure 6.1: Comparison of the different classifiers.

6.2 The combined classifier

The following section will present the accuracy of the combined classifier, in contrast to previously where each classifier was presented separately.

To be able to evaluate the combined classifier, an independent test set was retrieved. The gathering of data was performed in the same way as described in Chapter 4. The resulting test set contained 1795 accounts, where the number of females is 893 and the number of males is 902. Parts of the test set was utilized to train.

At this point, the models were only included if fundamental information was accessible, e.g. when a user did not provide a description, the combined classifier did not include the description model. In the new test set, the percentage of users providing a description was 72%.

When evaluating the mixed classifier, the accuracy was measured starting with only one of the models. Next step was to add another model to the mixed classifier. A classifier is only added if it conducts a higher accuracy. When the addition of the tweet style features model was added, it resulted in a loss of accuracy. Thus, this model is eliminated from the mixed classifier.

The tweet word features model achieved this time an accuracy of 74.7%. The description classifier gave the combined classifier an increase of 0.45%. Adding the username model resulted in an additional increase of 0.78%. Finally, the accuracy increased by 6.4% when adding the name model. The resulting accuracy of 82.35% was obtained.

6.2.1 An expanding combined classifier

In the following section, the confusion matrices for the combined classifier will be described. The measurements are further explained in section 3.3.

6.2.1.1 Starting point

At the very first, the combined classifier consists of the tweet word features model. The confusion matrix below, see Table 6.1, shows that 637 females were classified correctly, while 256 females were classified as male. Further on, 704 males were classified correctly, while 198 males were classified as female. Table 6.2 and Table 6.3 presents the performance measurements for the corresponding classes.

	Female	Male
Female	*637*	*256*
Male	*198*	*704*

Table 6.1: Confusion matrix for the tweet word features model

Females	
Precision	0.76
Recall	0.71
F1	0.74

Table 6.2: Performance metrics of females for the tweet word features model.

Males	
Precision	0.73
Recall	0.79
F1	0.76

Table 6.3: Performance metrics of males for the tweet word features model.

6.2.1.2 Insertion of description model

Continuing, the description classifier increased the accuracy by 0.45%. As can be seen in the confusion matrix below, see Table 6.4, the number of correctly classified females are 672, while 221 females are still wrongly classified as male. The number of correctly classified males are 677, while 225 males are still wrongly classified as female. This also shows that the amount of correctly classified men has decreased. Table 6.5 and Table 6.6 presents the performance measurements for the corresponding classes.

	Female	Male
Female	*672*	*221*
Male	*225*	*677*

Table 6.4: Confusion matrix for the combined classifier with the tweet word model and description model.

Females	
Precision	0.749
Recall	0.753
F1	0.751

Table 6.5: Performance metrics of females for combined classifier with the tweet word model and the description model.

Males	
Precision	0.754
Recall	0.750
F1	0.752

Table 6.6: Performance metrics of males for the combined classifier with the tweet word model and description model.

6.2.1.3 Insertion of username model

Continuing, the username classifier increased the accuracy by 6.4%. As can be seen in the confusion matrix below, see Table 6.7, the number of correctly classified females are 747, while 146 females are still wrongly classified as male. The number of correctly classified males are 616, while 286 males are still wrongly classified as male. This also shows that the amount of correctly classified men has decreased even more. Table 6.8 and Table 6.9 presents the performance measurements for the corresponding classes.

	Female	Male
Female	*747*	*146*
Male	*286*	*616*

Table 6.7: Confusion matrix for the combined classifier with the tweet word model, description model and username model.

Females	
Precision	0.723
Recall	0.837
F1	0.776

Table 6.8: Performance metrics of females for combined classifier with the tweet word model, the description model and username model.

Males	
Precision	0.808
Recall	0.683
F1	0.740

Table 6.9: Performance metrics of males for the combined classifier with the tweet word model, description model and username model.

6.2.1.4 The resulting combined classifier

The final step is to add the name model. The final accuracy reaches 82.35%. To use a combined classifier resulted in an increase of almost 7.4% compared to just using one classifier. Unfortunately, only females are classified more accurately. Table 6.11 and Table 6.12 presents the performance measurements for the corresponding classes.

	Female	Male
Female	*786*	*107*
Male	*210*	*693*

Table 6.10: Confusion matrix for the resulting combined classifier.

Females	
Precision	0.789
Recall	0.880
F1	0.832

Table 6.11: Performance metrics of females for the resulting combined classifier.

Males	
Precision	0.866
Recall	0.767
F1	0.813

Table 6.12: Performance metrics of males for the resulting combined classifier.

7

Discussion and future work

This chapter discusses parts of the result and tries to present risks to the validity in the used methods. The chapter also aims to give a comparison of the achieved result with previous research. Finally, we aim to highlight possible future extensions.

In this thesis work, we have extended gender inference on accounts that tweet primarily in Swedish. It has been clear that it is possible to infer the gender of Twitter users in Swedish contexts with the help of the mixed classifier. However, the lack of labelled data has limited the development further. The reason is that it is extremely time consuming to label the data manually. When discussing the manually labelled data, the accounts are manually classified by the author of this thesis. Even though we have tried to be careful when putting a label on the accounts, it is probable that some mistakes have been made. Thus, some accounts might be labelled wrong causing wrong classifications. Also, this small data set might make it difficult achieving stable results.

7.1 Comparison with previous research

The achieved solutions have to some degree been based on earlier research. The purpose of the following section is to put the achieved results in a bigger picture.

Burger et al. presents language-independent classifiers for predicting the gender of Twitter users [1]. Unfortunately, their sample of Twitter users is biased. They utilize Naïve Bayes, SVMs and Balanced Winnow2's, and compare the results. Their single most informative field was the user's full name which gave an accuracy of 89.1%. The final classifier achieves an accuracy of 92%.

Very limited work has been done when it comes to inferring gender in non-English contexts. Ciot et al. presents a study where similar accuracy can be achieved for languages differing from English using existing gender inference machinery [9]. They used a SVM-classifier presented by Zamal et al [3]. Additionally, they explore if unique features of languages other than English can increase accuracy. The study includes four languages: French, Japanese, Indonesian and Turkish.

Zamal et al. extends the existing work by looking at Twitter profiles and postings by friends [3]. The work includes predicting the gender, age and political orientation. They could conclude that inferences using only the features of a user's so called neighbors outperformed the ones based on the user's features alone. Worth noticing is that the dataset only included 400 users. The reason for this is due to the large amount of neighborhood data that is needed for every user. Gradient Boosted Decision Trees and SVMs were considered in the work. SVMs outperformed the other algorithm. Finally, there is no convincing improvement for gender inference by adopting the neighborhood data.

7.2 Future work

During the time of the thesis work, a number of interesting extensions to evaluate has been discovered. Due to time and scope limitations these have been neglected. This section aims to highlight some potential extensions.

First and foremost, a way to differentiate between company accounts and personal accounts need to be investigated. During the data collection phase, the data was manually inspected. This meant that company accounts were eliminated from the data set. In the future, we want the classifiers to handle all kinds of accounts.

Retweets have not been considered in this thesis. Thus, analyzing whether the retweeting tendency would present any further details about the gender would be of interest.

A limitation of the thesis work was the lack of labelled data. Additionally, the data was collected during a small period of time. But when finding a way to collect larger datasets, a possible and interesting feature would be to investigate when tweets are posted.

8
Conclusion

In this thesis, several methodologies for predicting gender of Swedish Twitter accounts were investigated. We have evaluated the capacity for existing inference methods to be used outside their intended English-language context.

To sum up, the main goal of the thesis work is reached, i.e to design and develop an automatic classification system that can infer users' gender of Swedish Twitter users. The thesis work included three subgoals. Two out of three goals are reached. Firstly, we have succeeded in analyzing whether tweets, username, the name of the user and the description is useful when differing between genders. Secondly, we have succeeded in constructing a mixed classifier in order to increase the accuracy. Unfortunately, analyzing the behavioral tweeting patterns of users was restricted by the dataset. The data retrieved was limited to a small period of time. Thus, the time of posting a tweet was not considered which in turn lead to issues when looking at, for example, tweeting sessions.

Bibliography

[1] J. Burger, J. Henderson, G. Kim, and G. Zarrella. *Discriminating Gender on Twitter.* Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, United Kingdom, 2011.

[2] J. Davis and M. Goadrich. The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd International Conference on Machine Learning.* pp. 233-240.

[3] F. Zamal, W. Liu, and D. Ruths. 2012. Homophily and latent attribute inference: inferring latent attributes of Twitter users from neighbors. In *Proceedings of the International Conference on Weblogs and Social Media.*

[4] W. Liu, F. Zamal, and D. Ruths. 2012. Using social media to infer gender composition from commuter populations. In *Proceedings of the When the City Meets the Citizen Workshop, the International Conference on Weblogs and Social Media.*

[5] M. Pennacchiotti, and A. Popescu. 2011. A machine learning approach to Twitter user classification. In *Proceedings of the International Conference on Weblogs and Social Media.*

[6] M. Zachary, B. Dickinson, and W. Hu. 2012. Gender prediction on Twitter using stream algorithms with N-gram character features. In *International Journal of Intelligence Science.*

[7] C. Manning, P. Raghavan, and H. Scütze. 2009. Text classification and Naive Bayes. In *An introduction to information retrieval*, Cambridge University Press, pp. 272-275.

[8] W. Liu, and D. Ruths. 2013. What's in a Name? Using First Names as Features for Gender Inference in Twitter. In *AAAI spring symposium: Analyzing microtext.*

[9] M. Ciot, M. Sonderegger, and D. Ruths. 2013. Gender Inference of Twitter Users in Non-English Contexts. In *EMNLP.*

[10] A. Munoz. 2014. Machine Learning and Optimization. Courant Institute of Mathematical Sciences.

[11] C. Manning, P. Raghavan, and H. Schütze. 2008. Support vector machines and machine learning on documents. In *Introduction to Information Retrieval*, Cambridge University Press.

[12] K. Murphy. 2006. Naive Bayes Classifiers, University of British Columbia.

[13] Z. Miller, B. Dickinson, and W. Hu. 2012. Gender Predicition on Twitter Using Stream Algorithms with N-Gram Character Features.