Author Profiling: Gender Identification

Juan Soler Company

MIIS MASTER THESIS UPF / 2013

THESIS SUPERVISOR
Leo Wanner Department TALN

Universitat Pompeu Fabra
Barcelona
To Susana, best friend and future mother of my children
Acknowledgements

I would like to thank a lot of people for this amazing experience that has been the first edition of the Master of Intelligent Interactive Systems. First of all, the university. If it wasn’t for UPF’s outstanding treatment, this experience wouldn’t’ve been the same. The professors, Hector, Leo, Ricardo, Vlad ... have taught us a lot and have treated us very well, so, I’m really thankfull for that. Specially I have to thank Leo, for being a great supervisor. Working with him has been so good that I decided to work with him for the next 3 years.

The other students have been my new family this year, we’ve worked together, and had lots of fun, becoming great friends. So, they have been another key part of this amazing year.

Of course I have to thank my parents, for always being there for me, supporting me, and basically always helping me in every possible way. My sisters, nephews, nieces, brothers in law... all need a little space in this section, cause they are part of my amazing family.

My friends need a bit of space as well, Massi, Mateu, Raul, Pedro, they all have been there for me and they are very important. Without them, my mental stability would be even more questionable than it is now.

To finish this section I have to thank Susana. If this year was a book, her face would be in the front cover. So, thanks for being born.
Abstract

Feature extraction has been used in lots of different scenarios. We can think of the feature extraction process that is done sometimes to characterize music or the one that we’ll be interested in, that is done on written texts. One of the applications of feature extraction in written texts is Author profiling: extracting as much information as possible from the author of these texts. We can see that this could be really interesting for commercial purposes and even in forensic scenarios. We’ll be working in Gender Identification, the goal of this work is to extract the gender of the author of written texts. Determining if the majority of your clients is male or female is very valuable information for big companies, so we’ll try to do this, with the best possible accuracy.
Contents

1 RELATED WORK 3

2 OUR APPROACH 9
  2.1 Theoretical Model ................................................. 9
  2.2 Features ............................................................. 10
    2.2.1 Character-based Features .................................... 11
    2.2.2 Word-based Features .......................................... 12
    2.2.3 Sentence-based Features ...................................... 13
    2.2.4 Dictionary-based Features .................................... 13
    2.2.5 Syntactic Features ............................................. 14
    2.2.6 Feature Summary ............................................... 14

3 EXPERIMENTS 17
  3.1 Dataset ............................................................. 17
  3.2 Experiment Setup .................................................. 18
  3.3 Results and discussion ............................................ 19

4 CONCLUSIONS AND FUTURE WORK 23
List of Figures

1.1 Results of [Koppel et al., 2000] .............................................. 4
1.2 Results of [Argamon et al., 2009] .......................................... 5
1.3 Results of [Cheng et al., 2009] .............................................. 6

2.1 Feature Summary ................................................................. 15

3.1 Dataset ............................................................................... 18
3.2 Results ............................................................................... 19
3.3 Results compared to baselines .............................................. 21
Introduction

The growth of Internet in the past 10-15 years has been huge. A direct consequence of that has been an explosion of data generation. Every day, people write in blogs, chat with each other, write e-mails... So, written texts are generated each second.

The thing about the net, is that everyone is as honest as they want to be, so, it’s fairly easy to lie about your gender, age, or even about your name.

Large corporations are always looking for ways of precisely know who their clients are. It’s very important for them to know if their clients are men or women, if they are young or old, or even the language they speak. When they know who their clients are, they can adapt their products to their target audience and as a result, earn more money.

Internet based companies such as Facebook, Google or Twitter have lots of data from their users, but these users can lie very easily, so, the obvious ways of knowing who their users are might not work very well. To get this information without distortion a good approach would be to cast this as an author profiling problem.

Author profiling is about extracting as much information as possible from the author of written texts. This information is extracted by analysing how the author writes: what words are used, what kinds of structures are used, the usage of certain characters... etc.

As these internet-based companies have lots of written material from their users, it’s easy to see that this approach can be quite successful. We’ve talked about Facebook and other big companies that were the more obvious cases, but another example could be a company like Amazon, which might want to see who are the people that review their products. We can see that we have lots of written material here as well, that can be used to profile their clients and as a result can give the company some information about what kind of products are more popular for what kind of people.

If we get one step deeper into the Internet we can also see lots of potential usages of this approach. In forums or even in pages like Reddit of 4chan, there are tons of written texts that users post. Lots of them only express themselves, but
there are some users that only want to make other people angry by insulting them or basically saying whatever they need to get under the other users’ skin. These kind of users are called trolls and they have to be managed, most of the times, manually, by having a moderator that can kick the user out of the system, or by having a flag system, in which other users can 'tag’ him as a troll. We can see that in this area, author identification could be useful as well. These users might have some underlying features in their writing that are common. If somebody detected those features, there could be a system that automatically detected trolls and for example, it could kick them out of the system without the need of having moderators or flag systems.

We don’t have to think about this field as an abstract field in which people theorize without having anything actually working. In our nowadays life, we have some examples. The most obvious one being of course the spam filter in our email accounts. These filters extract several features from the emails we receive, and they classify them according to some criteria that lets the non-spam emails (ham) go to our inbox, and moves the spam to the spam folder.

In this paper we will be getting into a specific case of Author profiling. This work is about extracting the gender of the author of written texts. In our case, the case of study will be blog posts. We’ll explain a fairly simple system that works well and that has some original ways of extracting features from the dataset.

It will be interesting to see what features distinguish male from female writing. It’s obvious that men and women are different. As we think in a very different way, it can be predicted that the way in which we express our thoughts will also be different.

In the following sections, the related work will be reviewed to know what has been done in this area. After that, we’ll explain our approach. We’ll start giving a brief theoretical model, so we are on the same page. Then, we’ll explain what features have been extracted and why. The next chapter will be about the experiments we made. In this section, the dataset that has been used will be explained and just after that we’ll see the experiment setup and the results our system got. Finally we’ll have some conclusions and future work.
Chapter 1

RELATED WORK

Extracting information from written texts to profile their author has been a topic of great interest in the last years. Several attempts have been quite successful. These attempts can be classified depending on what they try to identify and on what kind of dataset is used to do it. We’ll see some papers that try to accomplish the same goal we do and some that are more ambitious and try to identify not only gender, but also age and even native language at the same time. Others try to identify other features.

In the group of papers that try to accomplish the same goal we do but in another context, we have an interesting approach that was explained in [Koppel et al., 2000]. Their goal was the same we have: to predict the gender of the author of written texts. To do that, they used a corpus composed by 920 documents in British English that are labelled both for author gender and for genre (fiction, not fiction). Their approach consists on extracting several features to classify the texts. The number of features that they end up extracting goes up to 1081. As this is a fairly large number of features, a reduction of dimensionality is mandatory. The final number of features used after the dimensionality reduction is 128, a much more reasonable number.

The features that are extracted from the texts of their corpus are a list of 405 function words, a list of n-grams of parts of speech, punctuation marks, the 500 most common ordered triples, the 100 most common ordered pairs and all the single tags. They try to get the features that are as most topic-independent as possible. This part, as crucial as it is, is not described thoroughly in this paper. We can see that the features they select can be interesting if they are topic-independent.

Selecting that many features introduces a complexity that has to be reduced using a dimensionality reduction process. This is a key difference between this work and ours. Our number of features is much smaller, so this process is not required and the system is a bit less complex.

To classify the texts, a classifier that is a variation of the Exponential Gradient
Algorithm [Kivinen and Warmuth, 1995] is used. This algorithm is a generalization of the Balanced Winnow algorithm [Littlestone, 1988].

To test this system, they test each category of features by itself and they try the combinations that give the best accuracy. The conclusion at this point is that the best results are obtained when the function words and the PoS features are combined.

The results are shown in Figure 1.1. As we can see, they show the accuracies of the classifier in different subsets of their dataset. It would be interesting to see the accuracy of the system in the whole dataset, without having fiction and non-fiction separated. If the features are as topic-independent as the authors think, it should also work well mixing fiction and non-fiction texts.

Here we see the most different part between this work and our approach. They also classify the texts between fiction and non-fiction. In that context, their classifier works very well, and gets accuracy above 97%.

<table>
<thead>
<tr>
<th>Testing on Genre:</th>
<th># of docs</th>
<th>Train on All</th>
<th>Train on Fiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiction</td>
<td>264</td>
<td>74.5</td>
<td>79.5</td>
</tr>
<tr>
<td>Fiction / Female</td>
<td>132</td>
<td>74.8</td>
<td>81.7</td>
</tr>
<tr>
<td>Fiction / Male</td>
<td>132</td>
<td>74.2</td>
<td>77.3</td>
</tr>
<tr>
<td>Non-fiction</td>
<td>302</td>
<td>79.7</td>
<td>82.6</td>
</tr>
<tr>
<td>Non-fiction / Female</td>
<td>151</td>
<td>79.2</td>
<td>83.3</td>
</tr>
<tr>
<td>Non-fiction / Male</td>
<td>151</td>
<td>80.2</td>
<td>81.9</td>
</tr>
<tr>
<td>Arts (Non-academic)</td>
<td>16</td>
<td>76.0</td>
<td>76.3</td>
</tr>
<tr>
<td>Arts (Academic)</td>
<td>24</td>
<td>75.6</td>
<td>77.5</td>
</tr>
<tr>
<td>Belief &amp; Thought</td>
<td>24</td>
<td>85.0</td>
<td>85.0</td>
</tr>
<tr>
<td>Biography</td>
<td>54</td>
<td>87.0</td>
<td>90.0</td>
</tr>
<tr>
<td>Commerce</td>
<td>10</td>
<td>60.0</td>
<td>84.0</td>
</tr>
<tr>
<td>Leisure</td>
<td>16</td>
<td>85.7</td>
<td>81.3</td>
</tr>
<tr>
<td>Science</td>
<td>26</td>
<td>74.2</td>
<td>78.5</td>
</tr>
<tr>
<td>Social Science</td>
<td>22</td>
<td>77.5</td>
<td>83.0</td>
</tr>
<tr>
<td>Social Science (Academic)</td>
<td>38</td>
<td>82.9</td>
<td>78.4</td>
</tr>
<tr>
<td>World Affairs</td>
<td>42</td>
<td>79.2</td>
<td>82.9</td>
</tr>
</tbody>
</table>

Figure 1.1: Results of [Koppel et al., 2000]

If we look at more ambitious papers, we have to talk about [Argamon et al., 2009]. Here, the goal was to extract several predictions from their corpus. These predictions were gender, age, native language and personality of the author. We can see that the goal here is to profile the author of the texts.

For each of these classification problems, there is a different dataset used. Both style-related (function words and parts of speech) and content-related features are used. Some of these features are the 1000 words that appear frequently and discriminate best between the classes of interest (they determine this applying the information-gain measure), articles, conjunctions, prepositions, etc. It is not specified how many features are used, but the number seems to be high.
The corpus that is used to classify by gender, and age as well, consists of the full set of postings of 19320 blog authors that write in English. We will see that this dataset is much bigger than the one it was used in our work.

They study the features that are the most useful for gender discrimination, which are determiners and prepositions (male writing) and pronouns (female writing). In both classification problems, they achieve a good accuracy by using a classifier called Bayesian Multinomial Regression [Madigan et al., ].

To classify depending on the Native Language of the writer, a corpus called International Corpus of Learner English is used. The goal is to determine if the authors are Spanish, Bulgarian, Russian, French or Czech. To do this, their approach is to analyse the errors in these texts and establish a link between these errors and the characteristics of these languages.

Finally a corpus made of essays written by psychology undergraduates is used to determine if they are neurotic or not neurotic.

As we can see, the goals in this paper are very ambitious, and they tackle different problems all at once. It would’ve been interesting to test all these classifiers at the same time with the same corpus, to know the accuracy of all the classifiers combined.

The results of this work can be seen at Figure 1.2. We can see that the gender identification part has a nice accuracy in their dataset, that the Native Language part is the one that gets better results, and that the Neuroticism analysis is the one that has the worst results.

![Table](attachment:image.png)

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Style</th>
<th>Content</th>
<th>Style+Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (2 classes)</td>
<td>50.0</td>
<td>72.0</td>
<td>75.1</td>
<td>76.1</td>
</tr>
<tr>
<td>Age (3 classes)</td>
<td>42.7</td>
<td>66.9</td>
<td>75.5</td>
<td>77.7</td>
</tr>
<tr>
<td>Language (5 classes)</td>
<td>20.0</td>
<td>65.1</td>
<td>82.3</td>
<td>79.3</td>
</tr>
<tr>
<td>Neuroticism (2 classes)</td>
<td>50.0</td>
<td>65.7</td>
<td>53.0</td>
<td>63.1</td>
</tr>
</tbody>
</table>

Figure 1.2: Results of [Argamon et al., 2009]

Another paper that is related to the last one we talked about, and with our work is [Schler, 2005]. They use a blog post based dataset, which it’s similar to the one we compiled. In this paper we have a study on how gender and age influences the style of bloggers. It’s interesting because the context is similar to ours. They try to classify the blog posts by gender and age of the author. The dataset is a bit different than ours; theirs is composed of the posts of 71493 blogs of Blogger.

These posts can be formal or informal while our posts are fairly formal. The approach is similar to the ones that have already been discussed, features are ex-

---


2[http://www.blogger.com](http://www.blogger.com)
tracted and fed to a classifier. The features can be divided in style-based features (pronouns, assents, negations, determiners, hyperlinks, post length, etc.) and content-based features (words that appear more than 5000 times and have a high Information Gain coefficient). So we can see that the number of features is very high, and that there is an information gain process that is executed for each word in their dataset. We can see that this adds more complexity.

One of the interesting concepts that they point out is that prepositions and articles are used with increasing frequency as bloggers get older. The results they get in the gender identification scenario are about 80% of accuracy and approximately 75% in the age classification problem.

It’s interesting to note that author profiling has been done in many different scenarios. As it was mentioned, one of the contexts are blogs, but we can also see in the literature similar studies in the Email scenario [Cheng et al., 2009]. In this case, they try to identify the gender of the authors from a dataset called Enron email dataset. This corpus is composed by emails of employees of an energy company that went bankrupt and had their emails published. The corpus consists in 4947 male-written mails and 4023 female-written that have more than 50 words and less than 1000. We will see that this is a bigger dataset than the one we’re going to use.

In this paper, they extract features that are divided into five groups: character based features, word based features, syntactic based features, structure based features and function words. 68 psycho linguistic features are extracted using a text analysis tool called Linguistic Inquiry and Word Count (LIWC)\(^3\). The total number of features is 545, which is not as large as some other papers mentioned, but it’s still a pretty high number. Some of these features are the total number of lines, the number of interjection words, the number of colons, an entropy measure and number of characters per word to give some examples.

To classify the emails, two classifiers are used, Support Vector Machines (SVM) [Burges, 1998] with a radial basis kernel and decision tree [Rokach and Maimon, ]. The results can be seen in Figure 1.3.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>minimum words per e-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>73.38</td>
</tr>
<tr>
<td>SVM</td>
<td>80.08</td>
</tr>
</tbody>
</table>

Figure 1.3: Results of [Cheng et al., 2009]

We can see that the best performance of their system is achieved when facing mails that have between 100 and 200 words. It would be interesting to see what accuracy would this system achieve when facing the whole dataset.

\(^3\)http://www.liwc.net/
Another interesting group of papers are the ones that try to use social media to identify the gender of the author. [Burger et al., 2011] is an interesting example. In that case, the dataset is composed by the tweets of 184,000 users labelled by gender. 55% of the tweets of this dataset are written by women and 45% by men. The average number of tweets per user is 22. One of the interesting things in this work is that the tweets they use are written in different languages. To extract features from the tweets, the tweets themselves are used, as well as three fields from the Twitter user profile: Full name, screen name and description. These fields give extra information that the tweets might not have given otherwise. The features that are extracted are quite simple; each feature is a simple Boolean indicator that represents presence or absence of the word or character ngram in the set of text strings associated with the particular field.

The feature extraction process that is described in this paper is a complex process that requires over 20 gigabytes of storage and a feature pre-processing phase.

To do the classifying process, several classifiers were tested: SVM, Naïve Bayes [Lowd et al., ] and Balanced Winnow were some of them. Naïve Bayes achieved a 67% of accuracy, Winnow got 74% and SVM got 71.8%, but the training process required over 15 hours. When the tweets and the three profile fields are used, the accuracy goes up to a 92%, but in this process, some information of the users is used, so their written texts are not the only source of information.

In the scenario in which chat messages are used as dataset we have to talk about [Kucukyilmaz and Cambazoglu, 2006] and [Köse et al., 2008]. These two papers are attempts to extract the gender of the users in a chat scenario. Both try to predict the gender of the members of conversations (in Turkish) in different chat services. Some of the specific features that they use for their context are the smiley usage, abbreviations, slang words, and different function words. In both cases, the dataset is huge and they get accuracies, using different classifiers above 80%.

If we look at papers that try to identify different features we have to talk about [Koppel et al., 2005]. In this paper, they try to determine the author’s native language. In that case, a dataset made with the essays of foreign students writing in English is compiled. Their approach is to analyse the errors, and to use these errors to determine the native language of the author.

As all of the other mentioned papers do, in this one, a wide variety of features are extracted from their corpus. In this particular case, these features can be divided into three groups: Function words, Letter n-grams and the most important and decisive for their work, errors and idiosyncrasies. In the last group the idea is that the writers might transport orthographic or syntactic conventions from their native languages over to English in ways that result in non-conventional English. The focus of this paper is to describe which error types they look for, and
how they do that. The errors are categorized into four categories: Orthography, Syntax, Neologisms and Parts-of-Speech bigrams (rare PoS bigrams that show up). To get these errors the process they describe starts running a text through the MS-Word application and its embedded spelling and grammar checker. Then, the errors found are recorded along with the best suggestion.

The total number of features that are extracted is 1035, composed by 400 standard function words, 200 letter ngrams, 185 error types and 250 rare PoS bigrams. To classify the texts, SVM is used, and an accuracy of 80,2% is achieved.

If we look for papers that describe the problem the same way we do, we have to mention [Zhang and Zhang, 2010]. This is a gender prediction system that is developed in the blog scenario, which is the one we’re also going to use. The main difference is that they use blog posts from blog hosting sites like Blogger. So, the correctness of the posts is not guaranteed. In our case, the texts are guaranteed correct, and they express some sort of opinion. Another difference is that the dataset is bigger than ours, it’s composed by 3226 blog entries labelled by gender.

The biggest difference between this work and ours is the kind of features that are extracted. In their case, they keep track of the most common PoS tags, they also have 20 word lists (‘Family words’, ’at Home words’...) and finally they have a binary representation of the words and punctuations that appear (every word is a binary feature). The total count of features is 56024. As we can see, it’s immensely bigger than the number of features we use. There is a feature selection process that they need to do in order to get the features that matter the most and to turn this problem into a feasible one. Using Naïve Bayes, SVM and LDA to classify the text they get (best-case) an accuracy of 72,1%.

It’s interesting to see that nobody in the mentioned papers used a dependency parser to get the relations and dependencies between words of their dataset. This will be one of the innovations that are going to be implemented in our system.
Chapter 2

OUR APPROACH

2.1 Theoretical Model

The problem we’ve tackled was casted as a Supervised Learning problem. Our job will be to infer a function from our corpus that can be seen as labelled training data. Each one of our texts has been manually labelled using one of our two classes \{man, woman\}.

Using a subset of our corpus that will be called the training data, our system will infer a function that helps it determine the label of the testing data, which will be the rest of the dataset. The system won’t know the labels and it’ll have to predict them using the function.

Each document we have in our corpus will be represented as a n-dimensional vector. Each of these dimensions will be each one of the features we extract from the texts. This way, we have an abstraction of our dataset and we have the information we need to proceed.

The system will use these vectors to classify the texts into our two classes. To do that, we’ll be using a classifier, which is basically a function that uses some criteria to extract how alike the vectors are, and uses that to get a general way of classifying the new vectors which their class is unknown.

Lots of different classifiers are available; in our case the classifier that got us the best results was a Bagging approach [Bbeiman, 1996].

The output of our classifier will be the predictions it extracts, and using these predictions we can get the accuracy, which will be the percentage of the times our system works well.
2.2 Features

We’ve talked about how we need to get from pure text data, to a group of n-dimensional vectors that are composed by several features that we extracted. So, these features are the central part of our work.

The process of extracting these features was incremental. On a first step, the texts were analysed. We looked at the words that were the most used by men and women, we extracted the number of different words, the mean length of the texts, etc. Lots of features were considered, but in the end, only the ones that we thought that were relevant were used.

It will be interesting to see that the total number of features we use is quite small. We extracted a total of 83 features, which as we saw in the state of the art review is very small, if we compare it with other approaches in this field.

Our 83 features can be classified in five groups. There was a sixth group that was a parts-of-speech feature group, in which we analysed the grammatical category of each word, and used the number of pronouns, verbs, nouns, adverbs... we got on each text as a feature. This was discarded because it didn’t increase the accuracy of our system, so these features didn’t work well with our dataset.

The five groups of features we used are the following:

- Character-based
- Word-based
- Sentence-based
- Dictionary-based
- Syntactic

As we can see in these groups we analyse our texts from several different angles. We start with the smallest analysable part, the characters that compose our texts. Then we start analysing the words and sentences. These first three groups are the basic set of features we needed to have and are not a big innovation. The last two groups, on the other hand, are two approaches that in this field have not been used yet. We analyse words and sentences in a different way.

In the next sections of this paper, we’re going to explain what features compose each one of these groups, and we’re going to explain which were the hypotheses we wanted to prove with each one of these groups of features.
2.2.1 Character-based Features

The first group of features we have, analyses the text at a character level. We start by analysing the smaller units we have on our texts.

The hypothesis here is that men and women use characters in a different way. The upper cases, punctuation marks and other relevant characters might be used differently, so we try to look for these little details. We’ll see that this first group of features analyses very simple aspects of the text, but we’ll see in the results that this does not mean that these features are not relevant, sometimes simple approaches give better results than complex ones.

The features we extracted for this group are the following:

- Number of characters
- Percentage of upper cases
- Percentage of commas
- Percentage of dots
- Percentage of exclamation marks
- Percentage of question marks

First we have the number of characters that are used in the texts. Then it might be interesting to analyse the upper case ratio, a high number in these feature might tell us that the writer is using acronyms, or upper cases to express his/herself notoriously.

Then we get to the punctuation marks, which we thought that could be used differently. Commas for example, are a very personal way to divide your writing, so we thought that could be a differentiating factor between men and women. Each person has its own style, but the hypothesis was that these styles might fit in a pattern depending on the gender of the author.

Finally, the exclamation and question marks are analysed as well. This could tell us who is being more expressive, so it’s worth having it as features.

It’s interesting to see that we are using percentages. Given the fact that we have a quite irregular dataset, we have to do all the calculations in a way that the length of the text does not distort our results. In this case, the number of usages of each of the characters we’re looking for is divided by the total number of characters in the text.
2.2.2 Word-based Features

After looking at the texts at a character level, we have to analyse them using a bigger focus. In this case, we are going to look at the words that are used in the dataset. If we thought that men and women used characters differently, the next step is to see if they use words in a different way. We are going to look at the structure of these words as well as what kind of words are used.

The first step that was done in this area was to look at the words that were the most used in our dataset. We removed stop words, acronyms and proper nouns, and saw that in the most used words by men we had words such as:

- states, government, union, american ...

And in the women list we had:
- well, health, justice, children ...

So, we can see a pattern that will give us some clues about how women and men write, at least in our dataset. These clues will be used later in this paper (the dictionary based features use some of these clues as hypotheses).

In this set of features, we have:
- Mean number of characters per word
- Number of words
- Number of different words
- Percentage of stopwords
- Percentage of acronyms

We can see that we analyse the complexity of the words by counting the mean length of the words, this way, we can see if men or women use longer words than the other gender. Also, it’s interesting to look at the variety of usage of words, it’s reasonable to think that it might be a pattern regarding the richness of vocabulary in the texts.

The number of stop words and acronyms is also extracted as features. Basically in this group of features we’re trying to prove that the way men and women choose their words is different, the number of words used to express their opinion may also be different as well as the complexity of these chosen words. As we did in the last set of features, the number of stop words and acronyms are expressed as percentages. We divide the number of stop words and acronyms by the total number of words, so the length of the text does not distort our results.

In the section of the dictionary based features we’ll analyse words using a different approach.
2.2.3 Sentence-based Features

In this set of features, we analyse a bigger structure. We analyse the sentences we have on our texts. This group is directly related with the syntactic features that we’re going to see later.

In this group we only have two features:

- Mean number of words per sentence
- Number of sentences

As we can see those are very basic features. We count the number of sentences which can be seen as a way to see which gender is generating larger texts, which could be a good way to differentiate between male and female writing. Also the mean number of words per sentence is calculated. Using this feature we can see who is using more complex constructions.

2.2.4 Dictionary-based Features

This will be the first group of features that presents a big innovation. We’re going to analyse the words in our texts using a new approach.

Sentiment analysis is a very popular field nowadays, it’s used to determine the polarity of a text, basically if what is being said is positive or negative. It’s interesting to see that in the Gender Identification field, sentiment analysis techniques haven’t been used to classify. That’s one of the things that will be done in this set of features. Before getting into the details we’ll see the features we have in this group:

- Percentage of positive words
- Percentage of negative words
- Percentage of ‘patriotic’ words

In the first two features, we’re going to use a very simple technique used in sentiment analysis to extract features. Two polarity dictionaries are going to be used. These dictionaries contain words that are either positive or negative. These dictionaries have been developed in the university of Chicago\(^1\) and have more than 4000 words each. These features will be the number of positive and negative words we find in the texts divided by the total number of words.

When the texts were analysed a pattern was detected. It seemed that men were telling what was happening and women were thinking about the consequences.

\(^1\)http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
of these stories. So the main hypotheses here was that the emotional involvedness was a huge differentiating factor between men and women. Using these two features we’re trying to capture this emotional involvedness. As we saw on the results, women used positive and negative words in a much higher rate than men did, so in our dataset the hypotheses was confirmed.

Another thing that was observed was that men were more concerned about "patriotic" stories. The fact is that the blog posts we have are from an american blog, so we saw that patriotism had a lot of impact in the texts. To use this to our advantage, a new dictionary was compiled. The words that compose this dictionary are words that we considered 'patriotic', some examples of these words are:

- flag, patriot, homeland, independence ...

So the percentage of patriotic words in the texts was also used as a feature. We’re aware that this is a very context dependant feature, but as it worked fairly well and as it was the only one that depended on the context, we decided to keep it. To try the system on other contexts we might have to remove this feature or change the dictionary to adapt it to the new context.

2.2.5 Syntactic Features

The last set of features we have will be a new way of analysing the sentences. We’re going to focus on the structure of these sentences. To analyse the sentences we’re going to use a dependency parser, which will analyse each phrase and output the dependencies that we have between words in the sentences. The dependency parser that was used was Mate-Tools\(^2\).

What will be used as features will be the number of dependencies the parser gets of each of the possible dependencies that exist between words. These numbers will be expressed as percentages. The number of dependencies of each category is going to be divided by the total number of words on that text, so the length of the text does not distort the results. An example of feature in this group could be the percentage of subjects we have on a given text.

Apart from the usage of those dependencies, we’re also going to use as a feature the mean length of the dependencies. This is a way of analysing the complexity of the construction of the sentences. With this information we will see if there are important differences on how men and women build their texts.

2.2.6 Feature Summary

In Figure 2.1 we have a summary of the features that were used:

\(^2\)https://code.google.com/p/mate-tools/
We can see that the total number of features we’re using is 83 which is a relatively small number and a very small number if we compare it to the works we’ve analysed in the Related work section. We will see that this small number of features and the small dataset we use won’t be a problem to get a very nice accuracy.

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character-based</td>
<td>6</td>
</tr>
<tr>
<td>Word-based</td>
<td>5</td>
</tr>
<tr>
<td>Sentence-based</td>
<td>2</td>
</tr>
<tr>
<td>Dictionary-based</td>
<td>3</td>
</tr>
<tr>
<td>Syntactic</td>
<td>67</td>
</tr>
</tbody>
</table>

Figure 2.1: Feature Summary
Chapter 3

EXPERIMENTS

3.1 Dataset

The dataset that has been used is new and has been compiled for this work. The texts that we were looking for needed to be well written, in English and expressive. We thought that it would be very interesting to have expressive writing, so this way, we could use some sentiment analysis techniques to extract the gender of the author.

Several blogs were considered, but finally a blog in the NYTimes web page was chosen\(^1\). In this blog, several authors write about lots of different topics, that range from science to philosophy, also discussing current events, the economical situation worldwide and lots of other topics.

So with this dataset we have texts that are well written, in English, not always formal, and that discuss a wide range of topics.

The whole blog was crawled until late February 2013, so the dataset is composed by all of the blog posts that were written from the beginning of the blog to that date.

All of these texts were annotated. Three groups were found, the posts that were written by men, the ones by women and some texts that were written by more than one person.

In Figure 3.1 these groups can be seen in detail:

We can see that we have more than 5000 texts in our dataset, but also, we can see that the corpus is unbalanced. To be able to work with it properly we had to balance it.

The first step was to remove the texts written by more than one person, since they do not apply to our approach. Then, we selected all the texts written by women, and 836 texts written by men, so we have the same amount of texts written.

\(^1\)http://opinionator.blogs.nytimes.com/
One important thing to note is that these texts written by men that were selected were not selected using any criteria, so there’s no minimum length, no maximum length, not any restriction. This is a detail that we’ve seen in several papers, the dataset is filtered to match some minimum quality for the used texts and while that is a very respectable choice that maximizes the accuracy of the system, we’ve thought that it would be even more interesting to sacrifice a bit the final accuracy to have a realistic approach to this problem.

These texts are irregular, there are several that are only one sentence long and some that are a few pages long, so we can see that we have a fairly realistic and irregular dataset, which is not ideal, not perfect, and we can apply our system to a context that was not built to get a high accuracy but to have an approach that could be applied to real situations getting a good accuracy while having an irregular corpus.

### 3.2 Experiment Setup

To perform the described process, several technologies were used. In the feature extraction process, Python\(^2\) was used. The reason for this is the great potential this language has to process natural language, it has a whole library that was very useful to do that. This library is of course the Natural Language Toolkit (NLTK)\(^3\).

The features that were extracted were converted to dimensions of vectors that represented each one of the texts of the dataset, these vectors were written in an ARFF \(^4\) file that was fed to WEKA\(^5\), the program we used to do the classifying part.

To get reliable results we used a 10-fold cross validation process, so the results do not depend on which specific part was used to train and which part of the dataset was chosen to the testing.

---

\(^2\)http://www.python.org/
\(^3\)http://nltk.org/
\(^4\)http://www.cs.waikato.ac.nz/ml/weka/arff.html
\(^5\)http://www.cs.waikato.ac.nz/ml/weka/
The chosen classifier that was used in WEKA was bagging. AdaBoost [Freund and Schapire, 1997] also gave us decent results, but the bagging classifier worked even better.

### 3.3 Results and discussion

After the feature extraction process, the dataset was represented as a collection of vectors in which each of the dimensions was the value of the extracted features.

As we said in the previous section, we fed these vectors to a classifier and began the classifying process.

In Figure 3.2 we can see the results of our system. We can see for almost all the possible combinations of groups of features we have, the number of features that are used and the accuracy we get.

![Table of Results](image)

**Figure 3.2: Results**

In this figure we can see lots of information. First of all, we can see that with our system we get an 82,83% of accuracy which is excellent. The fact that we are showing the accuracy of the system using the whole dataset without any kind of restriction or without taking advantage of some feature of a specific part of the dataset makes it even better.
The usage of only 83 features with a small dataset makes this system very interesting given the relatively low complexity of it, and if we compare it with the works of the state of the art we talked about in the Related Work section we can see that this approach is not far away from the state of the art.

One of the concepts that was most arguable with was the inclusion of a context dependent feature in the system we designed. This was a feature that definitely would work well with this dataset, but would be completely useless if the context changed.

The great news is that if we watch closely the results in Figure 3.2 we can see that the best accuracy we get does not use the dictionary based features, so our system works at its best using only context independent features.

We can see that the decision of using a dependency parser was definitely a good one, using only the features of that group we get a 77.03%. With this set of features we’re analyzing how men and women construct their texts. This underlying structure of the texts is a very nice way of classifying mainly because it is context and content independent.

One of the other remarkable facts that we can extract from the figure is that we achieved a decent accuracy, 70.28% using only 14 features. It’s always interesting to try how good your system can be at the lowest possible cost. In some texts, all of the features we’re using might not apply, so using a subset of them (the ones that apply) can be a good idea, if that subset analyzes interesting features.

As we’ve seen, we have very good results, but it would be nice to compare them to some baseline to put them in context. We can use several baselines, the most basic one being a random classifier. Given the fact that we have two possible classes and the same amount of texts for each class, this classifier will have a 50% of accuracy.

If we look closely in the bibliography, we can see that most of the approaches use a set of basic features that are always present, these features usually are quite similar to our Character-based features and our Word-based features, and so these two groups of features can also be used as a valid baseline.

If we use these baselines, we can see in Figure 3.3 how our system performs compared to them:

Some very interesting things can be said about this graph. We can see that if we use our full set of features, we have an accuracy of 82.72%, which is a 32.72% higher than the random baseline and 16.27% higher than the W+C baseline. If we use the set of features that gets us the best accuracy, we have similar results. We can also see the accuracy of using only the syntactic features. As we said, this is a fairly innovative set of features, and seeing that it gets us 27.03% more accuracy than the random baseline and 10.58% more than the other baseline is definitely very good news.

We can sum up this section by saying that we used our features to determine
how accurate our system actually is. We’ve seen how well it works using all the possible combinations of the groups of features we extracted. A very good accuracy was achieved using only context independent features, and we also have to mention that with only 14 of the features we got a 70.03% of accuracy which was also a nice achievement.

We can see that the decisive factor of our system was the inclusion of the syntactic features; these features were the ones that made the system close to the results that are present in the state of the art nowadays.
Chapter 4

CONCLUSIONS AND FUTURE WORK

We’ve seen the whole development of this system step by step. We started by determining which features we wanted to use; we explained these features and also told which hypotheses we were trying to prove. After seeing the results, we can say that we were successful in our approach.

Some conclusions we extracted during this process were that there are important differences on how men and women write. These differences can be seen in a lot of different levels. We’ve analyzed them starting by reviewing the usage of characters, then we’ve looked at the words and sentences and we saw that there were important differences at those levels too.

One of our hypotheses was that the level of emotional involvedness was very different between both genders. We used two sentiment analysis dictionaries to measure this. Our hypothesis was that women were more emotionally involved when writing as opposed to men, that we thought that they tended to tell the stories without that emotional connection. After analyzing the texts we saw that this was (at least in this context) true, the number of positive and negative words that women used was higher than the ones men used.

We also saw differences in the usage of stop words, acronyms and in the usage of different words. It was very interesting to see that the number of words used by men and women was very similar, but if we counted the number of different words, that was not similar. In our dataset, the richness of vocabulary was definitely a differentiating factor as well.

The fact that the syntactic features worked so well tells us that there are important differences in how men and women construct their sentences. We can conclude that we found lots of differentiating factors to identify the gender of the author of written texts.

As far as future work, we have to say that this is only the beginning. We’re
definitely happy with our results, but this is only the start of a much complex system. It’s interesting to do Gender Identification, but we’re going to focus on identifying lots of other features from the authors of the text. The goal will be to have a system that profiles authors in a precise way. Apart from gender we’ll start focusing on age, native language, and even personality identification to mention some examples of features we might extract in the future.

Probably the first step will be to try the system that was discussed with other contexts, and other datasets to prove that this system can work in other scenarios. Another interesting approach might be to use texts that are written in different languages.

Right now we’re casting this problem as a supervised learning problem. It might be interesting to try a clustering approach to see how well it works.

To sum up, lots of different approaches can be taken from now on. The main goal will be to go from Gender Identification to Author profiling, which will be a more ambitious goal.
Bibliography


[Lowd et al., ] Lowd, D., Cs, P., and Edu, W. Naive Bayes Models for Probability Estimation.


