Gender Identification of Twitter Users

Maria-Stamatoula Karavolia
maria-stamatoula.karavolia@unifr.ch

Abstract

In this study, we present and compare various features for identifying the gender of Twitter users based on their tweets. Initially, we defined a coherent grouping of features combined with appropriate preprocessing steps for each group. The groups of features were the following ones: style-based features and the content-based features such as counts of Twitter hashtags or mentions and tf-idf n-grams of words. We formulate the gender identification task as a text classification task and we solve it using Support Vector Machines (SVM). The results for each of the features combined with SVM are compared thoroughly. The highest accuracy in determining the gender of the tweet author among all examined features reached 78.67% and was achieved by using the combination of tf-idf unigrams, bigrams and trigrams.

1 Introduction

With the rapid growth of web-based social networking technologies in recent years, gender identification problem has proven increasingly useful. Men and women are known to write in distinctly different ways, and these differences can be successfully used to make a gender prediction. Making use of these distinctions between male and female authors, we can automatically discriminate gender on labeled tweets from Twitter[1].

Many studies in gender identification have focused on the link between gender and language. Each gender exhibits different characteristic linguistic styles. Moreover, the length of tweets has several implications for gender identification. Because tweets are limited to 140 characters, there is less content available to predict an authors gender. On the other hand, the character limit for tweets means that users must fit whatever they want to say into a smaller space. This has the effect of concentrating the users writing style, increasing the necessity of using the characteristic text styles prevalent in social media. Studies have shown that women have a tendency to be more contextual and use involved language such as first and second person pronouns (you), present tense verbs (goes), emotions (love, angry, fear), certainty words (always, absolutely) when compared with men, and that men tend to be more formal and use informative language like prepositions (in, of, to), big words, nouns and numbers (million, three) [7].

In this study, we examine features which were used in the PAN 2015 [3] competition for solving the gender identification task and compare their performance on the training dataset provided by the organizers. Most of the submissions in the PAN 2015 competition use various features based on the word usage in user Tweets in combination with a Support Vector Machine (SVM) classifier. More specifically, in Section 2, we will provide an overview of related work. In Section 3 we will describe the problem of gender identification and highlight the goal of this study. In Section 4 we will describe our corpus, introduce the two categories of features, the classification method, the evaluation protocol and we will report our experimental results. Finally, in Section 5 we will conclude our paper providing future work directions.

2 State of the Art

Gender classification is one kind of text classification problem and there are several approaches that have been used previously to tackle the same task. For instance, in [4] regarding the gender classification task, they combined TF-IDF n-grams with style-based features. In [9] and [6] the authors studied the style of writing in blogs and for profiling the bloggers they used the learning algorithm

[1] http://www.twitter.com/.

Multi-Class Real Winnow (MCRW) to learn models that classify blogs according to author gender. The combination of stylistic and content features achieved the best classification accuracy. Similarly, in [5] and [2] the authors investigated gender identification from formal texts with accuracy 76% in tweet text in the second paper. Moreover, the author in [8] investigated how the style of writing is associated with personal attributes such as gender among others. The particular interest are findings that point to the psychological value of studying particlesparts of speech that include pronouns, articles, prepositions, conjunctives,and auxiliary verbs. In [10] experimented with short segments of blog post and obtained 72.1% accuracy for gender prediction. There are a lot of empirical studies devoted to blog mining or gender specific text analysis. [1] explores how gender affect writing style and topic using texts from blogosphere, and they found significant stylistic and content-based indicators.

3 Problem Definition

Text categorization is the process of classifying documents into a certain number of predefined categories. One of the problems in text classification is the identification of text genre which is becoming an important application in web information management. Gender identification problem can be treated as a binary classification problem i.e., given two classes male, female.

- $C = \{1, -1\}$ is a set of pre-defined categories for males and females.
- $D = \{d_1, ..., d_n\}$ is a set of documents written by users for which their gender is known.
- $Y = \{y_1, ..., y_n\}$ the set of target categories (gender) for each of the documents D.
  - When $y_i = 1$ the document was written by a female person.
  - When $y_i = -1$ the document was written by a male person.

The goal of gender identification is the classification of tweets into a fixed number of predefined categories $\{male. female\}$.

4 Classification Approach

For the gender identification task, a group of features combined with appropriate preprocessing steps for each group is proposed. In figure 1 the architecture of the system is displayed.

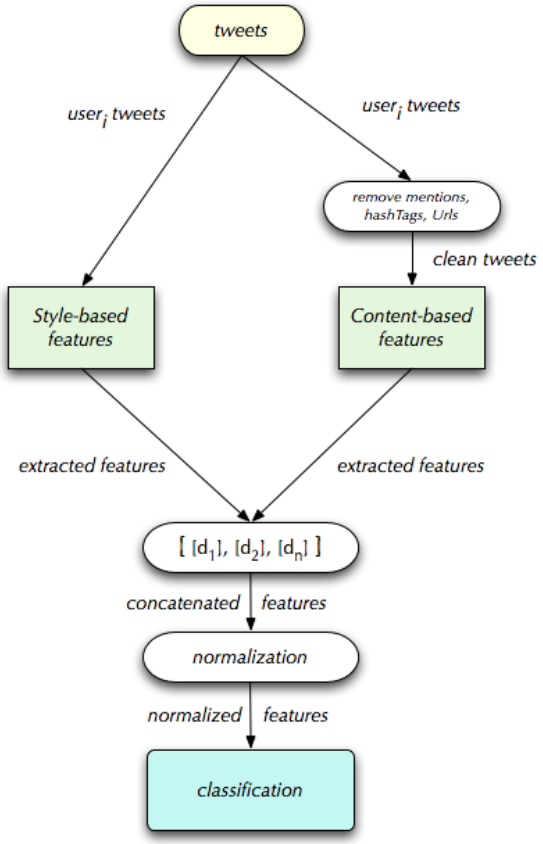


Figure 1: System architecture.

4.1 Features

Two different kinds of potential discriminative features are considered: style-related and content-related. Here, we examined four style-based features: counts of mentions, hashTags, Urls and punctuation signs features. As for the content-based features, we examined tf-idf of n-grams, bag of words n-grams, part of speech n-grams.

4.1.1 Preprocessing

Preprocessing is an important step as tweets contain specific information entangled in the text (hashtags, @replies and URL links). Therefore, a different preprocessing pipeline was applied to each group of features. There was no preprocessing done for style-based features. Content-based features preprocessing encompassed removing twitter bias such as @mentions, hashtags and URLs. To reduce the effect of Twitter bias, we removed the occurrences of '@username' in the tweets and we stripped of the character '#' from the hashtags.

4.1.2 Style-based features

The style-based features aimed to trace characteristics of the user text that are interdependent with the use of the Twitter platform. Here, these features include the counts of @mentions, hashtags and URLs in each users tweets.

4.1.3 Content-based features

The content-based group of features aim to capture characteristics of content that a user generates in spontaneous writing. Different features were tested, such as tf-idf of n-grams, bag of words n-grams, bag of words, tf-idf of words and part of speech n-grams. One of the features that used in experiments were the tf-idf n-grams.The tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (tf), i.e. the number of times a word appears in a document, divided by the total number of words in that document. The second term is the Inverse Document Frequency (idf), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears. More specifically the formula is defined for a term i in document j as follows:

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad (1)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Each user's document is thus represented as a vector of tf-idf n-grams. For the tf-idf unigrams 7112 unique tokens were resulted, while for tf-idf bigrams and trigrams 12.438 and 5637 unique tokens respectively. The tf-idf were created by using the Python library for topic modelling, document indexing and similarity retrieval, *Gensim*[2].

Moreover, a bag of words n-grams were created for our experiments and more particularly for each document a sparse vector of occurrence counts of words was created and that is, a sparse histogram over the vocabulary. The number of the unique tokens is the same as the tf-idf n-grams because the same corpus was used.

Last, a part of speech n-grams such as articles, nouns, pronouns, adjectives, verbs, prepositions, etc. were extracted using the *Stanford Log-linear Part-Of-Speech Tagger*[3] and then for each document a sparse vector of occurrence counts of words was created as before.

[2] https://radimrehurek.com/gensim/.

4.2 Classifier

Regarding the gender classification task, we used a Support Vector Machine (SVM) with a linear kernel. For the implementation of the above classifier, the *scikit-learn*[4] library is used. For the classification task the features were concatenated and were then normalized. Normalization was performed along instances so that each row has a unit norm.

5 Methodology & Evaluations

5.1 Corpus

In the gender identification experiments we chose to deal with tweets collected from the PAN-AP-2015 [3] corpus from Twitter in English. PAN, held as part of the CLEF conference is an evaluation lab on uncovering plagiarism, authorship, and social software misuse. Our corpus consists of 152 users each one of them have more than 20 tweets. The gender information was reported by the Twitter users themselves.

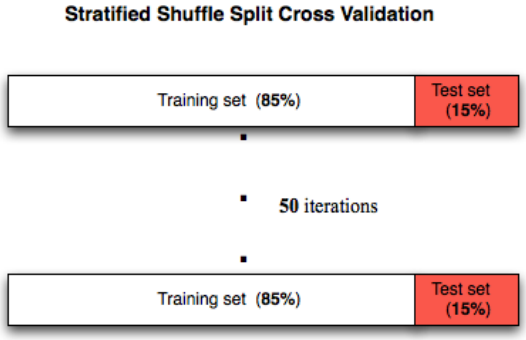
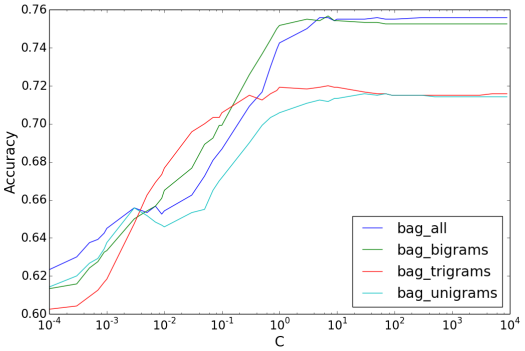


Figure 2: Stratified Cross-validation.

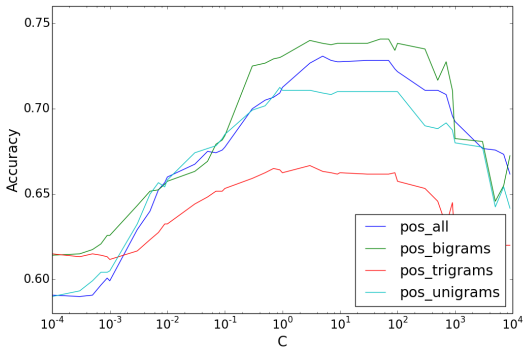
5.2 Performance Measures

For gender identification we used *stratified cross-validation* with shuffle split as evaluation protocol. In particular, we used the 85% for the training set and 15% for the test set (figure 2). This measure was used because the dataset was small and the training set was significant to have an equal number of patterns from each class (balance between the classes over the different the folds). We also defined the number of re-shuffling and splitting to

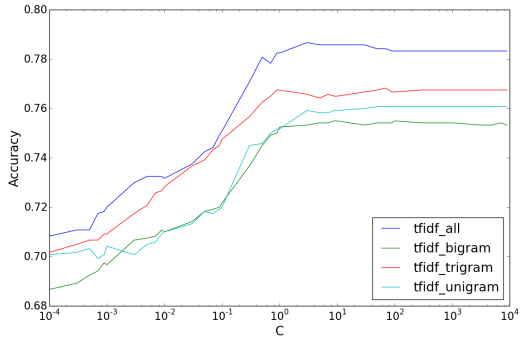
[3] http://nlp.stanford.edu/software/tagger.shtml
[4] http://scikit-learn.org/stable/



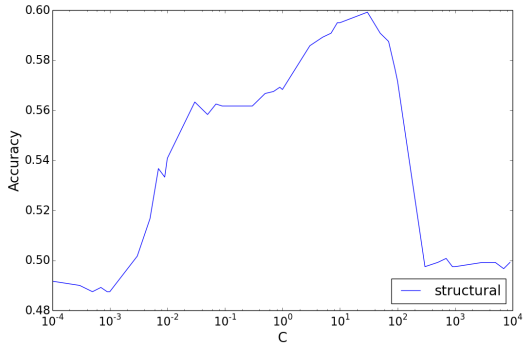
(a) Bag of words n-grams.



(b) Part of speech n-grams.



(c) Tf-idf n-grams.



(d) Structural features (mentions, Urls, hashtags and punctuation signs).

Figure 3: Features performance for different values of parameter C

50 iterations in order to obtain several folds from which we can get an accurate estimate of the performance.

5.3 Results

The results showed that SVM performed better when the content-based features were used. For a different values of the C hyper-parameter of the SVM, the accuracy of all the the feature combinations is displayed in Figure 3.

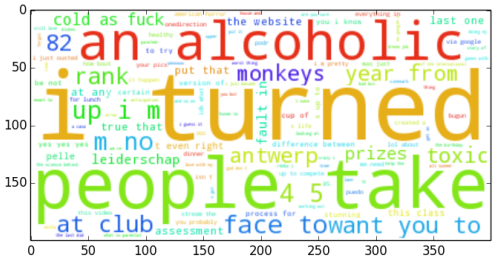
According to the C parameter value, the SVM decides the size of the margin hyperplane in order to avoid the misclassifying each training example. For instance, for large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. For very tiny values of C, you should get misclassified examples, often even if your training data is linearly separable.

The Table 1 shows the evaluation results for each group of features. We can see that the content-based features are more informative for the gender identification than the style-based features. Moreover, it can be also seen that most of the obtained accuracies are between 59.92% and 78.67%. Also, the results are more concentrated above 71%. are the combination of tf-idf unigrams, bigrams and trigrams with C=3, obtained the highest accuracy of 78.67%, while the style-based features achieved the lowest accuracy at 59.92%. Therefore, the style-based features are not useful for predicting the gender of users from his/her tweets.

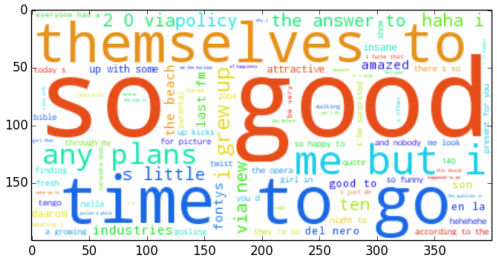
Interestingly, in the PAN 2015 competition the best performance was obtained by using the tf-idf trigrams, but in our case the combination of tf-idf unigrams, bigrams and trigrams achieved the highest accuracy.

5.4 Qualitative analysis

In this section we analyze qualitatively the coefficients learned by the SVM classifier. Here, we selected the feature with the highest evaluation accuracy, namely the tf-idf uni-



(a) Top 1000 useful terms for predicting male.



(b) Top 1000 useful terms for predicting female.

Figure 4: Two Word Clouds with including the tf-idf unigrams+bigrams+trigrams for male and female.

Table 1: Evaluation results in terms of accuracy for gender identification and on Twitter data.

Features	Accuracy
structural	59.92%
bag of unigrams	71.58%
bag of bigrams	75.67%
bag of trigrams	72%
bag of unigrams + bigrams + trigrams	75.58%
part of speech unigrams	71.25%
part of speech biigrams	74.08%
part of speech trigrams	66.67%
part of speech unigrams + bigrams + trigrams	73.08%
tf-idf unigrams	76.08%
tf-idf bigrams	75.50%
tf-idf trigrams	76.83%
tf-idf unigrams + bigrams + trigrams	**78.67%**

grams+bigrams+trigrams, and then, we visualized the terms in the feature space according to their coefficient learned by the SVM. More specifically, we first found the coefficients (weights) that SVM assigned to the features and then sorted them. The highest values belong to the positive class (female) and the lowest to negative class (male), so then we matched them to the terms in our dictionary.

Figure 4, displays the top 1000 terms that are the most strongest to predict if a tweet is written by a female or male: the higher the absolute value of the SVM coefficient the greater the size of the term. More specifically, in Figure 4a the most frequent terms that men use are **"i turned", "people take", "an alcoholic"**. Moreover, they make use of swear words like "cold as fuck" and numbers like "82" and "4 5". On the other hand, in Figure 4b the most frequent terms that women use are **"themselves to", "so good", "time to go", "me but i"**. Some other terms with lower frequency like "i hate that", "i grew up" which indicate negative emotion and also the use of personal pronouns. Lastly, in this particular dataset, we did not observe all the style characteristics which are used by males and females according to previous studies, for instance [7].

6 Conclusion

In this study, we found that content-based features are more informative according to identify the gender of a tweet, which confirms previous findings in the literature. More particularly, the combination of tf-idf n-grams offer the best accuracy at 78.67%. Conversely, the style-based features present the lowest accuracy at 59.92%.

Using this feature we presented the top 1000 terms that male and female are using most and we indicated that males are using more swear terms and numbers compared to women, that using more negative emotion and personal pronouns terms. However, using the tweets from the PAN 2015 dataset the discrimination between male and female based on their tweets is good (about 80%), but there is still much room for improvement.

Finally, one limitation of the examined approach is the curse of dimensionality. In our experiments, the dataset was relatively small but in the

case of bigger datasets the volume of the features space would increase significantly. Having a very high dimensional feature space can create computational complexity issues to certain classifiers which make more complex computations with the feature space than Support Vector Machines.

References

[1] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9), 2007.

[2] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[3] Francisco Rangel Paolo Rosso Efstathios Stamatatos, Martin Potthast and Benno Stein. Overview of the pan/clef 2015 evaluation lab.

[4] Krithara A. Giannakopoulos G. Grivas, A. Author profiling using stylometric and structural feature groupings-notebook for pan at clef 2015.

[5] Meyerhoff M. Holmes, J. The handbook of language and gender. 2003.

[6] Argamon S. Shimoni A.R. Koppel, M. Automatically categorizing written texts by author gender. In *Literary and Linguistic Computing 17(4). (2002).*

[7] Matthias R. Mehl and James W. Pennebaker. The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, 84(4):857–870, 2003.

[8] Mehl M.R. Niederhoffer K.G. Pennebaker, J.W. Psychological aspects of natural language use: Our words, our selves. page 547577, 2003.

[9] Koppel M. Argamon S. Pennebaker J.W. Schler, J. Effects of age and gender on blogging. *In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. AAAI (2006).*

[10] Zhang P. Zhang, C. Predicting gender from blog posts. Technical Report., 2010.