Importance of Unimportant Words for Authorship Identification

Pedro Filipe da Costa Dias Soares
pedro.dias.soares@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

April 2019

Abstract

There is an abundance of documents online and frequently this documents contain information that can be relevant for different applications. However, one of the problems associated with online documents is that frequently those documents are anonymous. Although identity cues are scarce in cyberspace, individuals often leave behind textual identity traces. Each author writes in a different way, thus by extracting the features from the text it is possible to identify the author of anonymous texts. It can also be used to determine if a text was written by the person claiming to have written it, or even to try and find the author of a given anonymous text. To correctly identify an author it is important not only to be able to correctly extract features from texts, but also to determine what are the features most suitable for the identification of the author. For our approach, we focused on the pattern of distribution of unimportant words, since we believed that each author has a specific distribution that will distinguish himself from any other.
Keywords: Unimportant Words, Portuguese Texts, Author Identification

1. Introduction

One of the major challenges of the document authorship identification problem is the extraction of the most appropriate features for representing the style of an author. Several techniques have been proposed, including attempts to quantify vocabulary richness, function word frequencies and part-of-speech frequencies [4]. This is because humans are creatures of habit, and as such have certain personal traits which tend to persist.

All humans have unique (or near unique) patterns of behavior. We therefore conjecture that certain characteristics pertaining to language, composition and writing, such as particular syntactic and structural layout traits, patterns of vocabulary usage, unusual language usage, stylistic and sub-stylistic features will remain relatively constant [4].

However, there still does not exist a consensus on the best possible set of features to determine a document's authorship. Unimportant words features are often disregarded, since it is considered to not provide any relevant information pertaining to the authorship of the text. However, we do believe that this feature has been misjudged, and that it may prove its usefulness when identifying the author of a text. As such, with our work we plan on determining the viability of the use of unimportant words as a feature in order to determine the authorship of a document. There has been a great number of studies published about the task of authorship identification of English documents. However, in comparison, the number of studies done for Portuguese documents can be considered marginal. As such, as the basis of our work, we will use Portuguese texts. With this approach we plan to determine the viability of not only unimportant words, but also other widely used features that provide very good results in English texts.

2. Background

This section presents previous works related with information extraction and author identification. These works addressed similar problems as the one we are facing now. These problems are nothing new, and as such by analyzing them we can better understand what might work better for our problem.

The current trend in author identification, is the identification and similarity detection in cyberspace. Here, the amount of texts published each day is of immeasurable quantity. However, the volume of texts, is not the only issue since the size as also proved itself to be a great challenge in the task of author identification. Many of texts that can be found are pertaining to conversations, or even posts in social media, in which the amount of words used is so small that it can be hard to extract any relevant feature.

Richmond et al. focused on the problem of the

identification of the author on blogs and e-books [7]. For their approach they used standard documents such as e-books to test the functionality of different algorithms, and then used them on online documents (blog entries and forum posts). Their datasets consisted of e-books, with 25 to 50 chapters, blog entries, each containing between 10 and 300 entries, and forum posts, each containing between 20 and 50 posts. First the documents where parsed in their software for feature extraction in which they extracted lexical and syntactic features like word counts, unique word counts and frequencies of function words. Then, they used a Bayesian classifier to classify the extracted features. Finally, they analyzed the results and concluded that using Naive Bayes and 62 text files having a total of 125865 words, they obtained an accuracy of 90.3%. As the number of texts reduced the accuracy also reduced, having obtained only 25% of accuracy when using just 4 texts in their dataset. From their results, they concluded that syntactic features such as common stop-function words and vocabulary are good features for large datasets and lexical features such as number of sentences and punctuation are good for small datasets.

Marcia Fissette [3] also studied how to identify the author of a short text. For her work, she relied on the structure of the text and the words that were used. Most of the features used for author identification are stylometric, especially in literary authorship. The data used for her approach was data extracted from a Dutch message board. From that data, there was a selection of 25 messages per author, and a total of 40 authors. The features used were: word unigram, word bigram, dependency triplets, word unigrams and dependency triplets, word bigrams and dependency triplets. To classify the results, Support Vector Machines where used, obtaining on average a success rate of 15.85% on retrieving the correct author.

Although identifying the author of small texts is of great importance, it is also important to determine what works best on texts of bigger dimension.

Most studies have determined the authors of texts using small pools of authors and texts. This gave a clear advantage to their approaches since as the number of authors increase, the number of unique characteristics of each author decreases. This also means that it is difficult to say if the techniques used where good or bad. Unfortunately, for the task of authorship attribution, the amount of people in this world creating content is huge, creating a large pool of possible authors. As such, it is important to determine the feasibility of identifying the author in large scale sets.
Arvind et al.[6] tackled this problem and for their approach they used a dataset comprising over 2.4 million posts taken from 100,000 blogs. In this dataset they extracted a set of features from each post and used them to train classifiers to recognize the writing style of each of the 100,000 blog authors. Most of the information from their dataset was obtained in ICWSM 2009 Spinn3r Blog Dataset, a large collection of blog posts. To clean their dataset, first they removed all html and any other markup or software-related debris, leaving only manually entered text. Next, they retained only those blogs with at least 7,500 characters of text across all their posts, or roughly eight paragraphs. Non-English language blogs were removed using the requirement that at least 15% of the words present must be among the top 50 English words. To avoid matching blog posts together based on a signature the author included, they removed any prefix or suffix found to be shared among at least three-fourths of the posts of a blog. Duplicated posts were also removed.
At the end of this process, their database contained a total of 2,443,808 blog posts, and an average of 24 posts per blog, where each post contained an average of 305 words, with a median of 223 words. The ten features that had the greatest information gain when computed were: the frequency of "", the number of characters, the frequency of words with only first letter uppercase, number of words, frequency of non phrases containing a personal pronoun, the frequency of full stops, the frequency of all lowercase words, the frequency of noun phrase containing a single proper noun, the frequency of all uppercase words and the frequency of commas.

They experimented with Nearest Neighbor (NN), Naive Bayes (NB) and Support Vector Machines (SVM), to determine what was the best classifier for this scenario. For each trial, they randomly selected three posts of one blog and set them aside as the testing data. The classifiers were then used to rank each blog according to its estimated likelihood of producing the test posts. They randomly selected three posts of one blog and set them aside as the testing data. The classifiers were then used to rank each blog according to its estimated likelihood of producing the test posts. They only selected blogs from the Spinn3r dataset as the source of test posts, but used the classifiers to rank all 100,000 blogs. In each trial, they recorded the rank of the correct blog. With this, they concluded that SVM's accuracy drops off rapidly as the number blogs increases. On the other hand NB and NN with a row norm normalization performed surprisingly well, obtaining the correct author in 8% of the cases.

It is important to notice that when trying to determine the author of a text, every bit of information can be useful. As such Maciej Eder [2] studied how much author information can be retrieved from

a raw string of characters in a text sample, without any kind of annotation, parsing, information retrieval, or keyword extracting. For this, the markers he chose were: The most frequent words, word bigrams, word tri-grams, word tetra-grams, letter bigrams, letter tetra-grams, letter penta-grams, letter hexa-grams, a combination of words and word bigrams and a combination of words and letter pentagrams. Then, the retrieved character strings were counted and the obtained numbers were converted to relative frequencies.
To make the results more reliable, a series of parallel attribution experiments were performed on corpora in four languages. The corpora were roughly similar containing seventy prose texts from 20 authors. The languages chosen for his texts were English, Latin, Polish and German.
To test his system he used Burrow's Delta platform. By using words and word n-grams as style markers he obtained 100% accuracy for English prose, when having a long vector of words (7500 from the top of frequency list) analyzed. As the number of words in the vector diminished so did the accuracy. When using 6-grams in English prose he obtained around 92% accuracy, and when combining stile markers he obtained an accuracy in the order of 95% for English prose.

N-grams is a feature that can be considered extremely useful in the task of authorship identification [1]. Vlado Keselj et al. [5] tested a theory of creating n-gram author level profiles, in order to determine the author of a particular text. For his experiment, his dataset consisted of two books for each author, in a total of three authors.
With this experiment, he obtained 100% accuracy when using unigrams and a profile size of the 20 more used n-grams. However, when the profile size increased the accuracy droped to 50%.
According to the information from this studies the best profile sizes are between 500 and 3000 n-grams, and the best n-gram size to create these profiles, are 5,6 and 7-gram. However, the amount of texts used to test this system, are low, making it hard to determine if the results that were obtained are trust worthy, and did not happen by luck.

On the other hand in some authors studies, the amount of texts used was of a greater size, making the information more reliable. John Houvardas et al. [4] used a dataset consisting of a training set containing 2500 texts from 50 authors, with 50 messages per author.

As features for classification they used the most frequently occurring character n-grams of variable length (3-grams, 4-grams and 5-grams). His proposed method for variable-length n-gram feature selection was to compare each n-gram with similar n-grams (either longer or shorter) and keep the dominant n-grams. Finally, a Support Vector Machine was trained using the reduced feature set, and obtained an accuracy of 73.08%.

3. Architecture

This section describes the architecture of the proposed approach for determining the author of texts extracted from a website containing technology related articles written in Portuguese. This approach is different from many others since, not only works with Portuguese texts instead of English texts, but also because it explores the possibility of using Unimportant Words as a feature in its' core. The work presented can be seen as an advancement over previous approaches, in the sense that it will focus on a language where classification algorithms are normally not tested upon, and because it tests the viability of a normally disregarded feature.

3.1. System Overview

Figure 1 illustrates the ideology of the developed system, namely the systems' input, processing components and output. As shown, the system receives as input, a set of training and test documents. First, it extracts the features of both training and test documents received as input. Then, the features extracted from the training documents are used to train our classifier. Finally, in the evaluation module, the previously trained classifier will receive the features from the testing documents, and will return his best guess regarding the author of the evaluated test document.

In order to determine what features work best in the task of authorship identification, we have used not only unimportant words features, but also features mentioned in our related work. These features include for instance N-Grams, Sentence Length and Number of Words. The flexibility of the developed system allows the user to select any combination of features providing the user the ability of determining which combination yields better results. The use of features that are not Unimportant Words, helps us compare the results obtained by these features, with the results obtained from the more commonly used features.

3.2. System Details

In this section we will present in greater detail each of the different components of our solution. These are divided into three main components, Inputs, System Components and Outputs.

The inputs component is composed by a Dataset, the List of Features, the List of Classifiers and the Number of Folds:

- Dataset: Our dataset is composed by 600 texts written in Portuguese. These documents were extracted from the website pplware.sapo.pt

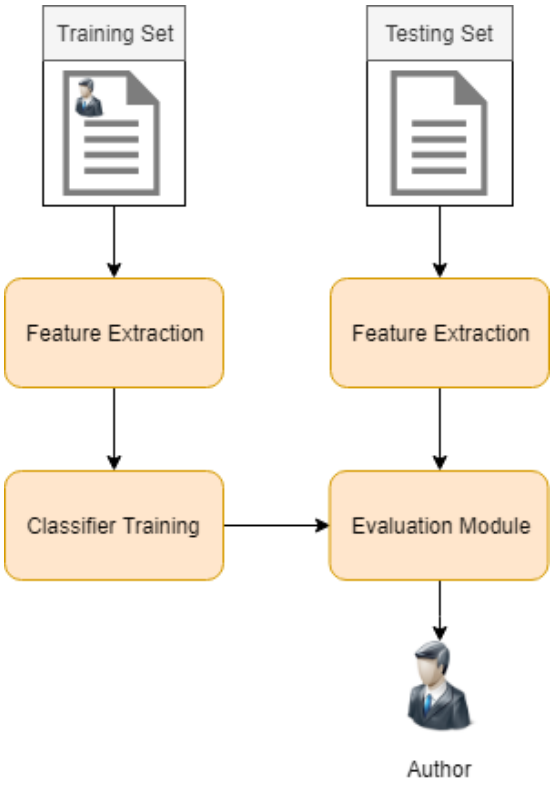

Figure 1: System Overview

and the extracted collection consist of 100 documents from each of the six different authors. The subjects and the length of the documents varies arbitrarily, although the general topic of each of the documents is related with technology. Each of the documents extracted were manually processed in order to remove all images, author names and html tags with the objective of leaving only plain text files.

- List of Features: Based on the previous related work, we concluded that for our approach the features that could best help us in the task of authorship identification were the following:

  N-Grams (One-Gram, Two-Gram, Three-Gram, Four-Gram and Five-Gram), Unimportant Words, Position of the First Unimportant Word, Position of the Last Unimportant Word, Percentage of Upper Characters, Percentage of Special Characters, Amount of Numbers in a Text, Average Length of Each Sentence, Number of Sentences of Each Text, Number of Words of Each Text, Average Length of Each Word, Percentage of Sentences that end with Full Stop, Percentage of Sentences that end with Question Marks, Percentage of Sentences that end with Exclamation Marks, Percentage of Sentences that end with Question Marks, Percentage of Sentences that end with Etc.

- List of Classifiers: In order to determine which classifier is best suited for the task of author identification for the features available, we tested our approach with several of them. The classifiers made available on the developed system are: Naive-Bayes, Support Vector Machines, Decision-Trees, Nearest Neighbours and Back Propagation.

- Number of Folds: In order to test the system, we used 10-fold cross validation. By using 10-fold cross validation we ensure that the results obtained were not due to a coincidence in choosing the training and text sets, since K-fold cross validation determines the mean of the results obtained in each of the test sets. However, the flexibility of the system, allows the user to select any number of desired folds through the user interface.

The System Components is composed by the Segmentation Module, the Feature Extraction Module, the Training Module and the Evaluation Module:

- Segmentation Module: When our system receives the dataset, it will segment the information into sentences and tokens, which will in turn be used for extracting the desired features, that will be used by the feature extraction module.

- Feature Extraction Module: In this module the segmented text will be used in order to extract each of the features selected by the user. After the features have been extracted, a text document will be created containing values for each of the desired features, and their respective author. These documents will then be written in a Weka framework readable format.

- Training Module: This module receives a Weka training document written by the Feature Extraction Module. The document will be divided into training and test sets. For this we will use the 10-fold cross validation method. Thus, the system will be trained with the training section, while the test section will be used later on, to evaluate the trained classifier.

- Evaluation Module: In this module each classifier will be evaluated 10 times (due to the 10-fold cross validation), using the previously created classifier against the respective test sets. The results of each of these evaluations will then be recorded in order to be presented on the Results Report, containing the mean of its 10 iterations.

Lastly the Output of the system is the Results Report.

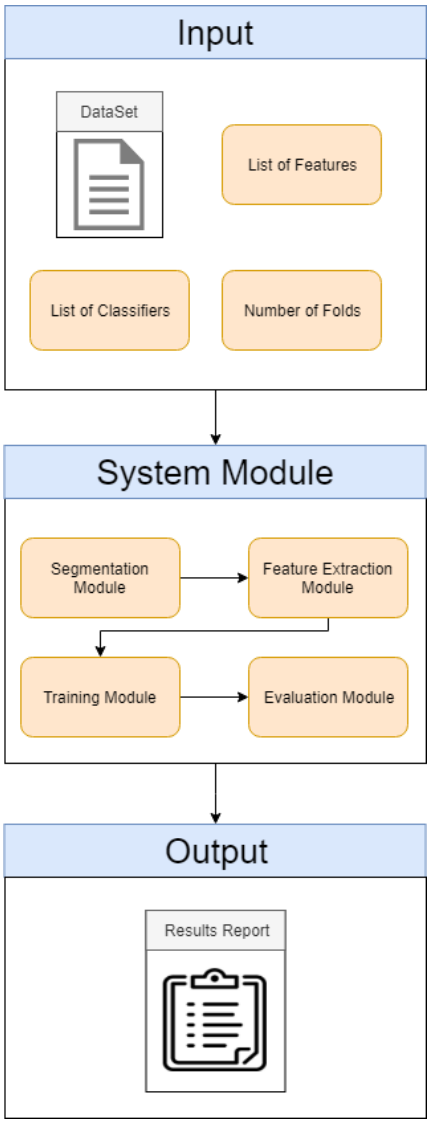

Figure 2: System Details

- Results Report: After the evaluation segment has run, a results report will be shown to the user. This report will contain the results obtained by each classifier upon using the selected features. The results report will be a text file containing the name of the classifier, the success rate and the features used.

In a nutshell, in terms of process, the system receives a set of documents which are parsed, and all images, author signatures and html tags are removed. The resulting documents will become our dataset. Next, the Segmentation Module will split each document of the dataset into tokens.

Afterwards, the Feature Extraction Module uses the tokens from the dataset, to extract the features selected by the user. Then, the system splits the dataset into 10 parts of the same size being one of the groups a test set, and the remaining will be the training set. The features from the training set will then be used by the Training Module, in order to train the selected classifiers.

Finally, the Evaluation Module uses the features of the test set, to try to guess the authors of the test set. The result is recorded and this process is repeated 9 times, one for each fold. Afterwards the result returned will be the average of all the results obtained for each of the 10 folds.

3.3. Experimental Tests

For our approach, the main feature we focused on, is the distribution of unimportant words. Although this feature is normally removed since it is believed that it provides very low to no information regarding the author, we believe it may be enough to identify him, or at least help in the task of author identification, in combination with other features.

Since there is not a consensus regarding the best feature or group of features in the analyzed related works, we decided to include several of them in this work, as a way to compare its' results with our suggested Non Important Words features.

Ideally, all possible combination of the presented features should be tested in order to determine which combination achieves the best result in the tasks of author identification.

However, since testing all the possible combinations and provide a meaningful analysis is impractical, and since the aim of this work is to determine how our proposed feature compares in relation with the most used features, we decided to group the feature into related themes namely: N-Grams, Unimportant Words, Document Content, Type of Document, Type of Narrator. The goal of grouping the features is to test and analyze all the possible combinations of features within each group and then to compare the results with the results of the best combination of features of the remaining groups.

Finally, the objective is to test all the possible combinations of the best features of each group and perform the same kind of analysis. The content of each group of features is as follows:

- Group 1: N-Grams - This groups is based of only N-Gram features. In our approach we consider only N-Grams of size 1 to 5, since bigger N-Grams revealed no significant results in previous related works.

- Group 2: Unimportant Words - This group of features is based on our approach in determining the validity of features related with often disregarded unimportant words. For this group we consider the percentage of occurrences in a text of Unimportant Words, the position of the first Unimportant Word and the position of the last Unimportant Word.

- Group 3: Document Content - The purpose of this group of features is to determine if the content of text that is being analyzed contains any author signature. For this group we determine the percentage of upper characters, the percentage of special characters and lastly the amount of numbers in a text.
- Group 4: Type of Document - This group of features seeks to determine the type of document that the author tends to write. If the author tends to write short texts with long sentences, or if he tends to write long texts but with short sentences and so on. As such, for this feature group we evaluate the average length of each sentence, the number of sentences of each text, the number of words of each text and the average length of each word.
- Group 5: Type of Narrator - This set of features intends to determine the type of the narrator that wrote the text. It tries to determine if the author writes in an assertive way, if he makes questions, and so on. As such, for this group the features that we have considered are: the Percentage of Sentences that end with Full Stop, the Percentage of Sentences that end Exclamation Marks, and the Percentage of Sentences that end with Etc.

In order to evaluate our approach, we compared our proposed feature, Unimportant Words, against groups of features from previous similar approaches in this field.

To do so, the classifiers used for our approach were Decision Tree, Naive Bayes, Support Vector Machines, Back Propagation and Nearest Neighbors. Each feature and feature combination was run through every classifier, and for each document the classifier returns the name of the author it believes is the correct one. Each answer is classified as correct or incorrect, then an average of correct answers is calculated, and the feature or feature combination that displayed the best average is considered the best for each of the classifiers.

The main objective of this is to determine which "Unimportant Words" features are the most relevant, and if these features would be part in the best combination of groups of features, or if their presence would just led to obtaining worst results.

4. Experimental Results

This section has the objective of describing the experimental results obtained, as well as their respective analysis with the intent of validate the viability of the proposed approach in the task of author identification. Each feature was tested individually in order to determine which features provide the best results for each classifier. Then, the features were tested in their respective groups by combining them with all other features of the group and determining which combination of features performed best. Finally, we once again expanded our approach, by testing all combinations of the best group of features from each group.

4.1. Experiments with individual features

When testing each feature individually, we came to the conclusion that in average the best results were obtained by the Decision Tree classifier. However, this classifier was not much superior than the others since Back Propagation and Nearest Neighbors performed almost as well. On the other hand, Support Vector Machines classifier under performed when in comparison with the remainder of the classifiers. In general, we can conclude that for almost all classifiers the 4-Gram feature was the one that obtained the best results in the task of author identification.

It is worth mentioning that although none of the features regarding unimportant words achieved great results, they all obtained results better than random chance and that the best classifier varies with the feature itself.

4.2. Experiments with Group 1 features

Based on our experiments with the features of the N-Gram group, we can conclude that overall the feature that obtained the best results was 4-Grams. However, all features performed similarly, being the combination of 1, 4 and 5 grams the one that obtained worst results. The best performing classifiers was Decision Tree followed by Back Propagation, and the worst performing classifier was Naive Bayes, although there was not a great variation between the classifiers.

4.3. Experiments with Group 2 features

The experiments performed with the Unimportant Words features, led us to the conclusion that the combination of features that obtained the best results were the Position of the First Unimportant Word, with the Position of the Last Unimportant Word. All the features performed similarly and the feature that performed worst was the combination of Number of Unimportant Words with the Position of the First Unimportant Word. The best performing classifier was Nearest Neighbours and with the exception of Support Vector Machines, all classifiers performed on pair.

4.4. Experiments with Group 3 features

Overall, for the group of features related with Document Content, the combination of features that obtained the best results was the Percentage of Special Characters with the Amount of Numbers in a Text. All features performed similarly, however when combining features, all feature combinations had a good increase in accuracy in comparison

with their performance alone. The classifier that obtained best results was Back Propagation, and the worst was Support Vector Machines. Support Vector Machines classifier distinguished from the remaining classifiers by its low performance, even though for all the features tested, it still obtained better results than the random chance.

4.5. Experiments with group 4 features

Based on our experiments with Type of Document features, we can conclude that the best feature combination in the task of author identification is the combination of Number of Sentences of each Text with the Number of Words of each Text and the Average Length of each Word. This feature combination did not obtain clearly better results than others, although there were some feature combinations that did obtain clearly worst results when in comparison with this feature combination. The feature that obtained the worst results in this test was Average Length of each Word, obtaining in some instances results worst than random chance. The best and worst classifier changed according to the number of features used. While for one feature the best classifier was Decision Tree, for two and three features was Back Propagation. It is also worth mentioning that overall, the best performing feature combinations, were combinations of three features, instead of a bigger combination of features.

4.6. Experiments with group 5 features

Based on our experiments with Type of Narrator Features, we can conclude that the best feature combination in the task of author identification is the combination between Percentage of Sentences that end with Full Stop, Percentage of Sentences that end with Exclamation Mark, Percentage of Sentences that end with Question Marks and Percentage of Sentences that end with Etc. This feature combination did not obtain clearly better results than others, with the exception of the Percentage of Sentences that end with Full Stop feature, which was the worst feature on the test, obtaining in some instances results worse than random chance. The best classifier changed according to the number of features used. However, independently of the test, the best classifiers were Decision Tree classifier and Back Propagation classifier, being the Decision Tree classifier more predominant. It is also worth mentioning that although, the best performing feature combinations was the combination of Percentage of Sentences that end with Full Stop, Percentage of Sentences that end with Exclamation Mark, Percentage of Sentences that end with Question Marks and Percentage of Sentences that end with Etc, the results did not improve in all cases where extra features were added.

4.7. Experiments with combinations of groups of features

The performance of each group of features is an important milestone for our work. However, the objective of our work is to determine the usefulness of Unimportant Words features, and to determine if the presence of these features may improve the results of other more used groups of features. We have already presented the evaluation and results of each group of features. In this section the features that obtained the best results of each group will be combined in order to determine what combination of groups of features performs well together and achieves the best results in the task of author identification. In a nutshell the best feature combination of each group of features are:

- Group 1: N-Grams - 4-Gram
- Group 2: Unimportant Words - Position of First Unimportant Word and Position of Last Unimportant Word
- Group 3: Document Content - Percentage of Special Characters and Amount of Numbers in Text
- Group 4: Type of Document - Position of Last Unimportant Word, Number of Words of each Text and Average Length of each Sentence
- Group 5: Type of Narrator - Percentage of Sentences that end with Full Stop, Percentage of Sentences that end with Exclamation Mark, Percentage of Sentences that end with Question Mark, Percentage of Sentences that end with Etc

All possible combinations of the mentioned groups were tested, and based on our experiments with combination of feature groups, we can conclude that the best feature group combination is Group 1 with Group 2 and Group 3. This combination contains the following features:

- 4-Grams
- Position of First Unimportant Word
- Position of Last Unimportant Word
- Percentage of Special Characters
- Amount of Numbers in a Text

This feature group combination did not obtain clearly better results than others, although there were some feature group combinations that did obtain clearly worst results when in comparison with this combination. The worst feature group combination was Group 3, although when in combination

with other feature groups, it did prove to be an important feature group.

The best and worst classifier changed according to the number of groups of features used. However, in average the best one was Back Propagation classifier, followed by the Decision Tree classifier. The classifier that obtained worst results was the Support Vector Machines classifier, although there was not a significant difference in the results between the best and the worst classifier.

With this experiment we can conclude that although the Unimportant Words group did not obtain very good results on its own, when in combination with other groups, it provided positive results. We can also conclude that since the Unimportant Words group, is part of the group of features that obtained the best results, this group of features can play an important role in the task of author identification. As such, although alone this group of features seems useless, it should not be disregarded. Thus, making it an important feature to be used in future works, when trying to determine the author of a text.

5. Conclusions

In this work, we have thrived to determine the author of Portuguese texts. In order to do so, we have presented a set of new features that are normally disregarded, namely Unimportant Words. The idea behind this set of features, is that since each author writes in is own particular and unique way, he is bound to have a unique style of writing Unimportant Words. As such, each document will have a unique signature of this features, and this can be used in order to better identify the author. Due to the impracticability of testing all the possible combinations of typically used features plus the ones we proposed, we created groups of features based on their similarity. We evaluated each group in order to determine which features produced better results alone, and which combination of features in each group would provide the best results. Finally, we combined the groups of features to try and understand if the Unimportant Words feature group, would have a positive or negative impact on the task of author authentication, when in combination with other feature groups. Through our tests we could observe that alone, Unimportant Words features obtained a barely above random chance of identifying correctly the author of a text. However, when this feature was used in combination with others, it increased the chances of correctly identifying the author of a text, and thus proving its usefulness on the task of identifying the author in Portuguese texts. We can also conclude, that for our experiments, the classifiers that obtained the best results were Nearest Neighbors and Back Propagation, and the one that performed worst was Support Vector Machines.

Despite the good results obtained, there is also a great amount of work to be done. Firstly, more tests should be done with similar texts, in order to validate the results obtained, and to make sure these results were not due to a coincidence.

These performed tests should be applied to a bigger corpus of texts and authors, in order to confirm the obtained results.

These features should be applied to other fields and languages, in order to determine if the results are not exclusive to the Portuguese language.

All feature combinations should be tested, and not only the best features of each group. It is also necessary to expand the amount of features used when testing.

The unimportant words list should also be expanded to include other non important words, as well as create a similar list for other languages.

Just like Unimportant Words features were disregarded, there may be several other disregarded features, that may achieve positive results when tested in different languages. As such, it is necessary to test these features in each language, in order to determine their importance in the task of author identification.

Acknowledgements

I would like to say a few words to all those that have helped me on this journey.

Firstly, I would like to thank my parents Abel Soares and Patrocinia Soares and to my brother Sergio Soares for everything they regularly do for me and for making my life a lot easier. Not only did they help me constantly over the years, they gave me opportunity to focus on my work, and provided me with the motivation needed during the complicated moments.

Secondly, I would like to express my most sincere gratitude to professor Andreas Miroslaus Wichert, for his availability, his precious advice, for supervising my work, for giving me great support and always believing in me. He showed great patience and always motivated me during this journey. Antonio Varela, who taught me the required knowledge for my degree and my future work.

My colleagues, Renato Nunes, Miguel Viola, Francisco Bom, Luis Rodrigues and Fernando Macedo, for constantly motivating me, and helping me solve the problems I faced.

Finally, to all my teachers who granted me the bases to complete this dissertation.

References

[1] M. Corney, A. Anderson, G. Mohay, and O. De Vel. Identifying the authors of suspect e-mail. *Communications of The ACM - CACM*, 01 2001.

[2] M. Cristani, G. Roffo, C. Segalin, L. Bazzani, A. Vinciarelli, and V. Murino. Conversationally-inspired stylometric features for authorship attribution in instant messaging. pages 1121–1124, 2012.

[3] M. Fissette. Author identification in short texts. 2010.

[4] J. Houvardas and E. Stamatatos. N-gram feature selection for authorship identification. *AIMSA*, 2006.

[5] V. Ke Selj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. *Proceedings of the Conference Pacific Association for Computational Linguistics PACLING'03: 2003*, 09 2003.

[6] E. Stamatatos. Author identification using imbalanced and limited training texts. *18th International Workshop on Database and Expert Systems Applications (DEXA 2007)*, pages 237–241, 2007.

[7] R. H. R. Tan and F. S. Tsai. Authorship identification for online text. *International Conference on Cyberworlds*, 2010.