The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Content-Dependent Versus Content-Independent Features for Gender and Age Range Identification in Different Types of Texts

M. Zakaria Kurdi
Department of Computer Science, University of Lynchburg

Abstract

This paper is about the comparison of content-dependent and content-independent features for the identification of short texts author's age range and gender. Eight content-dependent features based on profiles of ngrams of words are used. In addition, ninety-eight content-independent features covering all the linguistic aspects of texts from phonology to discourse are used. These features were extracted from three corpora of different sizes and types. Experiments were conducted using four different machine learning algorithms combined with these features. The results show that content-dependent features do a better job for gender identification on the three corpora. However, content-independent features did better with the task of age range identification.

Introduction

We are currently witnessing the growth of usage of electronic media in different forms of communication such as emails, blogs, and electronic newspapers. Attributing a gender and age range to the author of an electronic text is sometimes crucial for detecting fraudulent activities.

Many features can be used to identify the gender and age range of a text's author. Some are content or vocabulary-dependent and therefore require application-specific data. Examples of such features are sequences of actual words and characters. For the same author, these sequences may change depending on so many factors such as the social context (familiar interlocutor vs. unfamiliar), the topic (e.g. sports event, a family issue, or a natural phenomenon). This makes these features less desirable. On the other hand, there are generic features that depend on the language style rather than on the vocabulary. Although these features are not completely independent from the type of text or message, they are nevertheless much less dependent than the features of the first type. Hence, these features can be used with a plug and play mode on virtually any dataset without prior training on the application data. Most of the previous works in the literature relied solely on content-dependent features or on a mixed set of content-dependent and content-independent features. This makes it hard estimate the role played by each group in the classification process.

In this paper, two different sets of features are explored, one is content-independent and another which is content-dependent. Three different corpora, with three different types, are used to test if the relevance of these features depends on the data type. In addition to this comparison, unlike many previous works who focused on a limited number of features, this work uses ninety-eight features covering all the linguistic levels from phonology to discourse. Besides, experiments were conducted with four different machine learning algorithms.

State of the Art

Given the importance of author's gender and age range identification in many application fields, this subject attracted the attention of researchers from disciplines such as linguistics, psychology, and Natural Language Processing.

For example, Newman et al. (2008) conducted research on 1400 text samples, which resulted in the conclusion that there are significant differences between women's and men's texts. They found that women use more words that are related to emotions and social relations, while men's vocabulary is more objective and impersonal.

Using n-grams and functional words as classification features, Argamon et al. (2003) proposed a text classification method for authors' gender attribution. Their work combined stylometric and classification techniques to achieve an accuracy of about 80% in an author's gender identification in the general case. When the text genre is defined, the accuracy goes up to 98%. On the other hand, Coyotl-Morales et al., (2006) proposed a method that characterizes documents by a set of word sequences that combines functional and content words by means of a process for mining frequent word sequences. They reported 76.8% of accuracy on a corpus of 353 poems.

Sarawgi et al. (2011) conducted a study on two corpora of scientific texts and web blogs about the author's gender

identification with features that are supposed to be both topic and genre neutral. They found that the best approach is based on character-level language models that learn morphological patterns. The main limit of this study is that it uses a few linguistic features (eight lexical and syntactic features). Besides, one can argue that the sequences of characters are just a translation of words or sequences of short words, which makes them a content-dependent feature.

Shrestha et al. (2016) conducted research on a large corpus of 85000 users of the DailyStrength health support forum. Their work aimed to detect a user's age and gender from their forum posts. They used a mixture of generic features like word and character sequences and familial terms as well as features that are forum specific such as users' names. Two methods are adopted to classify the ages. The first uses five age ranges, which cover about 10 years each, and the second uses three age ranges: 13-17, 23-27 and 33-42. They reported 61.23% successful classifications on age prediction, with five age groups and 65.39% with three age groups. No experiments with content-independent features were reported.

Corpora

Three different corpora are used in this study. This combination is motivated by the diversity of the nature of the considered corpora in terms of their sizes, the covered topics and the number of authors. This makes it possible to compare different types of approaches.

Enron[1] emails is a large corpus that is freely available. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's scandal. It is made of about 500.000 emails (1.32 GB of raw data) from 150 authors. To get the author's gender, the first name of the sender is extracted from the sender's email. It is then compared against the lists of the male and female names available from NLTK. If the name is only available in one of these lists, the author is then added to the list of considered authors. This filters unisex names like Ashley or Alex. After cleaning, only the body of the email sent by the source author is kept without any history or attached documents. Furthermore, only emails with 30 words or more are kept. As a result, the used subset contains only 149454 emails with 76790 emails written by males and 72664 written by females. The subjects of the emails are not limited to professional interactions as they cover some social activities outside the work environment. Given its large size, this corpus covers authors from different professional ranks and different social and ethnic backgrounds. Nevertheless, the high education level of the authors and the readers of the emails is the main common point.

An extract of the Reuters' C50 newspapers articles corpus[2] is used. It is made of 2200 texts written by 22 professional authors, eleven of whom are females. This makes this corpus gender balanced. The genders of the authors are determined manually based on their names. Being professional writers, the style of each of these authors may be different. However, given the journalistic nature of these texts, one cannot expect extreme variations of style, such as very complex sentences syntactically. The clarity being the priority in such texts. The average size of a text extract is 291 KB. All the articles use 30 words or more.

The third corpus is about blogs[3]. The language used in these blogs is usually informal or semi-formal, and the subjects are diverse (e.g. daily activities, movies, and political events). A total of 14786 blogs is available in the corpus, after filtering the blogs with less than 30 words. 14646 blogs were obtained with 8192 blogs written by males and 6454 written by females. The authors' ages range between 13 and 47 years. The average size of a blog is about 1.11 KB. The authors genders and ages are provided with the data.

Feature Extraction

As seen in the literature review, a wide variety of linguistic and nonlinguistic features have been used to identify the gender and age range of authors. In this paper, two sets of features are used. A pool of ninety-eight content-independent features that cover all the areas of linguistic complexity such as phonology, morphology, lexicon, syntax, and discourse are used (Kurdi, 2017b). What makes these features content-independent is that they do not require training on the specific domain of the text to be effective.

Phonology provides an abstract description of the sound structure of the language in terms of both segmental level (phonemes or syllables) and supra-segmental level such as stress and intonation (see (Kurdi, 2016) section 2.1.2 for an introduction to these issues). Phonology is hence an important descriptor of both spoken and written language. Three phonological features are considered. These features are the mean numbers of graphemes, phonemes, and syllables per word. In addition, seven writing formulas are calculated for each document. The considered formulas are the Gunning's Fog index, the Flesch-Kincaid formula, the Coleman-Liau Index, the Spache readability formula, the

[1] https://www.kaggle.com/wcukierski/enron-email-dataset

[2] https://archive.ics.uci.edu/ml/datasets/Reuter_50_50#

[3] https://www.kaggle.com/rtatman/blog-authorship-corpus

Dale–Chall formula, the Automated Readability Index (ARI), and the FORCAST Readability Formula. Many of these formulas combine phonological complexity measured by the number of character or syllables per word with lexical complexity.

Morphology is about the study of the form of the words as a function of their linguistic role. The considered morphological features include features related to word complexity such as the mean number of prefixes and suffixes per word, the diversity of stems (word morphological roots), as well as thirteen verb tenses. Verb tenses are detected with a module that uses hand-crafted regular expressions of POS tags sequences to recognize different verbal constructions. Finally, the percentages of thirteen POS tags in the text (like nouns, simple adverbs, and comparative adverbs) are also counted as features.

Lexicon is considered from the angle of words' meaning as well. It is commonly admitted that the text's meaning is the result of the combination of the meanings of the individual words that are constituting it. Hence, fourteen lexical features are examined here as possible candidates for classifying the texts by complexity. Several lexical measures are considered like lexical density, lexical sophistication (Read, 2000), (Hyltenstam, 1988). Lexical diversity with features like Type Token Ratio (TTR) and its two main corrections that were proposed to solve its bias toward the size of the text: Guiraud's corrected TTR (GTTR) and Caroll's corrected TTR (CTTR) are also considered. Furthermore, to account for the lexical sophistication, the Verb Sophistication Measure (VSM) (Harley and King, 1989) is calculated. Practically, are considered as sophisticated the verbs whose frequency rank is higher than 200[4] in the McMillan English Dictionary, which contains a list of the 330 most frequent verbs[5]. To find the uninflected form of a verb, the verb conjugation module, provided within the Pattern.en toolbox[6], is used.

Syntax is a key indicator of style. To extract the syntactic features, two freely available toolkits are used. For parsing, the Stanford Parser[7] is adopted. Some functionalities from NLTK[8] such as the sentence tokenizer and the POS tagger are also used.

Thirty-seven syntactic features are used. They cover the different aspects of sentence syntax. For example, the mean number of phrases and the mean length of phrases cover the extent of a sentence, which is a source of diversity; some use long sentences while others use shorter ones. On the other hand, the percentage of inverted declarative sentences[9], the number of subordination per sentence, the mean phrase coordination per phrase, the mean height of parse trees, and the percentage of complex T-units per T-units are all used to measure different aspects of syntactic styles. For instance, some authors prefer simple sentences connected with conjunctions, while others favor complex sentences.

Ngrams of POS tags are also considered. The idea here is that the more diversified the sequences of tags, the richer are the used syntactic structures in the text. Two measures are used: ngrams (bigrams, trigrams, and fourgrams) per sentence and ngrams per words, where the number of the ngrams is divided by the number of sentences and number of words respectively.

Furthermore, three ngram models (bigrams, trigrams, and fourgrams) for all the texts are built and used to enrich their author's gender profiles: three ngram profiles for male authors and another set of three ngram profiles for female authors. After building the author's gender profiles, the three ngram models of individual texts are used to calculate the distances between them and the genders profiles of their type. This gives a Zipfian distribution of the ngrams. For example, the distances between the bigram model of a text and the male and female bigram profiles are calculated. The obtained distances are used as features. In theory, if the author is a female, the distance between the bigram model of the text and the female bigram profile should be smaller than the one with the male bigram profile. The distance between a text ngram model and a gender ngram profile is the absolute value of the subtraction between the ranks of the ngrams in the two models.

Discourse connectors yield an important insight into the complexity of information structuring within the text. This factor is calculated as the number of lexical connectors divided by the number of words within the text. Another feature is also based on the same principle but with argumentative discourse connectors only.

Using the textblob[10] library's sentiment analysis functionality, the sentiment of every text is calculated and used as a feature. As a sentiment can be positive or negative (within the range -1, +1), an absolute value is also used. This helps to test if men or women use more emotional language regardless of the polarity of the emotion.

The coherence of the text is measured using the distance between the nouns of every pair of contiguous sentences. This helps estimate the overlap of the ideas. Informally, this is done by calculating the mean Wu-Palmer

[4] 200 being an empirically defined threshold.
[5] http://www.acme2k.co.uk/acme/3star%20verbs.htm
[6] http://www.clips.ua.ac.be/pages/pattern-en
[7] https://stanfordnlp.github.io/CoreNLP/
[8] https://www.nltk.org/
[9] Where the subject follows the tensed verb or modal.
[10] https://textblob.readthedocs.io/en/dev/

similarity[11] (WUP) distance between every noun in the current sentence and all the nouns in the previous sentence. For more details, see equation 1 about the coherence calculation of a sentence and equation 2 about the measurement of the distance between a word and a sentence.

$$coherence(S_m) = \frac{\sum_{i=1}^{n} sim(N_i,\ S_{m-1})}{n} \text{[eq.1]}$$

Where n is the number of nouns in the sentence S_m and N_i *is a noun from* S_m.

$$sim(N,S) = \frac{\sum_{i=1}^{k} dist(N,N_i)}{n} \text{ [eq. 2]}$$

Where k is the number of nouns in the sentence S, N_i is a noun from S, and N is a noun from the next sentence. The function dist is the WUP similarity. Finally, the mean of the similarities of all the sentences of the text is calculated and used as a coherence score of the text.

Eight content-dependent features are used. They are based on four word ngram profiles (unigrams, bigrams, trigrams, and fourgrams) and the comparison of text profiles with male and female profiles in the way described above for the sequences of POS tags. The distance between a text model and the corresponding male and female profiles is calculated.

Age Range Identification Experiments

Predicting the age range of an author depends on many complex features such as its intellectual maturity, language mastery, and culture. Unfortunately, these features do not depend on the age only. Other factors, such as the degree of education and social context, play a key role in determining the style of a text. For example, a highly educated young author may have a richer vocabulary than an older one with lower education. Furthermore, an author with a rich language may simplify his expressions when addressing people, he thinks have lower level of language abilities. This makes age range predication a difficult task.

To reduce this problem, two different approaches to consider age ranges are adopted: a discrete approach and a gradual one. In the first approach, there are only differences in writing between adults and teens: 13-17 and 18+. The main issue here is that in the blog corpus, the only used corpus that is labeled with ages, there are 9405 texts by authors who are above 18 and 3456 ones who are under 18.

According to the second model, people's writing evolves throughout their life. Hence, there should be tangible differences between writers for every decade in life. This gives four classes for the existing age ranges within the blog corpus: 13-17 with 3456 texts, 20-29 with 5411 texts, 30-39 with 3821 texts, and 40-47 with 173 texts.

In all the following experiments, four different machine learning algorithms were used, Random Forest[12] (RF), AdaBoost[13] (AB), Neural Nets (NN) and Logistic Regression (LR). These algorithms are chosen for their higher performance after preliminary experiments with a wide range of algorithms that included SVM, Decision Trees, and Naïve Bayes. The Orange datamining toolkit[14] was used in these experiments.

Four performance measures are reported: Area Under the Curve (AUC)[15], Recall, Precision and F1 (please refer to (Kurdi, 2017a) for a detailed introduction to these measures). Note that AUC is designed for binary classification that is why AUC results are more accurate for Model 2. Cross-validation is used with 20 folds.

Age Range Identification Results

The results of the age classification on the blog corpus with content-dependent and content-independent features are presented in tables 1 and 2 respectively.

Discussion of Age Prediction Results

As seen in table 1, content-dependent features are not very effective in determining the age range of the author (the result range is close to those reported by (Shrestha *et al.*, 2016)). A possible reason to that is that authors from different age ranges use similar vocabulary patterns. Table 2 shows that content-independent features are more effective in predicting the age, with Neural Nets being the most effective in the binary and continuous scenarios. With both content-dependent and content-independent features, the performance is clearly higher with the binary approach to model the age. In addition to the fact that binary classification is naturally easier than classification with a higher number of classes, the corpus imbalance could be an extra reason for difficulty.

[11] WUP similarity between two words is a score based on the depth of the two senses these words in the taxonomy and that of their most specific ancestor node.

[12] https://en.wikipedia.org/wiki/Random_forest

[13] https://en.wikipedia.org/wiki/AdaBoost

[14] https://orange.biolab.si/

[15] https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve

	Classifier	AUC	F1	Precision	Recall
binary	RF	**0.65**	**0.70**	**0.69**	**0.72**
	AB	0.60	0.48	0.48	0.48
	NN	0.71	0.52	0.53	0.54
	LR	0.58	0.64	0.66	**0.72**
Contin.	RF	0.71	**0.53**	**0.53**	**0.54**
	AB	0.60	0.47	0.47	0.47
	NN	0.71	0.52	**0.53**	**0.54**
	LR	0.60	0.36	0.41	0.42

Table 1. Results of age range identification on the blog corpus with content-dependent features

	Classifier	AUC	F1	Precision	Recall
binary	RF	0.83	0.79	0.79	0.80
	AB	0.66	0.73	0.73	0.73
	NN	**0.87**	**0.82**	**0.82**	**0.83**
	LR	0.65	0.73	0.63	0.71
Contin.	RF	0.74	0.56	0.56	0.56
	AB	0.60	0.47	0.47	0.47
	NN	**0.78**	**0.60**	**0.60**	**0.60**
	LR	0.61	0.36	0.44	0.44

Table 2. Results of age range identification on the blog corpus with content-independent features

Gender Identification Results

As seen in the literature review section, previous works have shown that males and females use different features of language, especially when it comes to vocabulary. The goal of this section is to determine how vocabulary-based features compare to general linguistic features in predicting the gender of a text's author. The question here is if the content-dependent and content-independent features play similar roles as in the age range identification. The experiments are conducted on the three corpora to see if factors like the domain of application or the size of the corpus play a role. The results of the experiments on the blog, Reuters's C50, and Enron corpora are reported in tables 3, 4, and 5 respectively.

	Classifier	AUC	F1	Prec.	Recall
Content -indep.	RF	0.72	0.66	0.66	0.66
	AB	0.58	0.58	0.58	0.58
	NN	**0.76**	**0.66**	**0.66**	**0.66**
	LR	0.63	0.59	0.59	0.59
Content -dep.	RF	**0.99**	0.97	0.97	0.97
	AB	0.96	0.96	0.96	0.96
	NN	**0.99**	**0.98**	**0.98**	**0.98**
	LR	**0.99**	**0.98**	**0.98**	**0.98**

Table 3. Results of gender identification on the Blog corpus

	Classifier	AUC	F1	Prec.	Recall
Content indep.	RF	0.81	0.73	0.73	0.73
	AB	0.63	0.63	0.63	0.63
	NN	**0.84**	**0.76**	**0.76**	**0.76**
	LR	0.64	0.60	0.60	0.60
Content -dep.	RF	**1**	0.99	0.99	0.99
	AB	0.99	0.99	0.99	0.99
	NN	**1**	**1**	**1**	**1**
	LR	**1**	**1**	**1**	**1**

Table 4. Results of gender identification on the Reuters corpus

	Classifier	AUC	F1	Prec.	Recall
Content -indep.	RF	0.95	0.88	0.88	0.88
	AB	0.91	**0.91**	**0.91**	**0.91**
	NN	0.94	0.85	0.85	0.85
	LR	**0.93**	0.85	0.85	0.85
Content -dep.	RF	**0.99**	**0.99**	**0.99**	**0.99**
	AB	**0.99**	**0.99**	**0.99**	**0.99**
	NN	**0.99**	**0.99**	**0.99**	**0.99**
	LR	**0.99**	0.98	0.98	0.98

Table 5. Results of gender identification on the Enron corpus

Discussion of Gender Identification Results

Gender identification with ngrams of words with the three corpora is either perfect or near perfect with all the used ML algorithms. The reason for this high performance is that the adopted content-dependent features turn the problem into linearly separable with the three used ngrams (see figure 1 for an example with trigrams). As seen in table 3, the best classifier, which is NN, with the content-independent features on the blog corpus has an F1 of 0.66 and an AUC of 0.76. NN are known to be effective classifiers of both small and large datasets. This limited performance can be explained by the relatively small size of the blog corpus as well as the diversity of topics and social backgrounds of the authors and the readers.

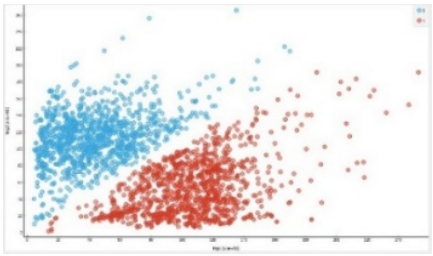

Figure 1. Scattered plot of the distance between trigram male and female profiles in the Reuters corpus

NN outperforms the other classifiers with the content-independent features on the Reuters corpus as well, with an overall mid-range performance. There are several possible

reasons for this outcome. The corpus is perfectly balanced. In addition, there are some professional and cultural similarities between the authors (they are all journalists with high educational and intellectual levels) who are targeting the same audience (news readers).

The results on the Enron corpus are clearly higher than the two other corpora. With the Enron Corpus, AdaBoost provides the best result, as it is known to perform well on data with high dimensions. In this case, the gap between content-dependent and content-independent features is bridged. This could be the result of two factors: the large size of the corpus and the limited social context of the interactions, which makes the stylistic differences more salient.

Conclusion and Perspectives

In this paper, a comparison was conducted between eight content-dependent word ngram features and ninety-eight content-independent features for text's authors age range and gender identification. Three different corpora of different data types and sizes were used in the experiments. The results show that, despite their limited number, content-dependent features provide perfect or near perfect classification with the three used corpora. This is because they turn the problem into a linearly separable one. These features do a poor job in identifying the authors' age range, however. Conversely, content-independent features do a better job for age range identification. As for authors' genders identification, their results vary from borderline to good, depending on the used corpus.

There are multiple areas of improvement that can be covered in a future work. A thorough examination of the role of each feature in identifying both the gender and age can be conducted. Furthermore, this study can benefit from adding more corpora of different types such as literary, and scientific corpora. Larger corpora labeled with authors' age can also bring more insights. Despite the good size of the used feature set, some more features can be added, especially to cover more aspects of discourse structure. Ngrams of characters can also be considered to enrich the pool of content-dependent features. The informally conducted experiment shows that features from all linguistic levels play a role in obtaining the presented classification results. Experimenting with different feature selection methods can help achieve an optimal balance between the outcome and the size of the feature set. Finally, a minimum of 30 words is used as a threshold for texts. More experiments with different thresholds can be conducted.

Bibliography

Argamon, S.; Koppel M.; Fine J. and Shimoni, A. R. 2003. Gender, Genre, and Writing Style in Formal Written Texts, Genre, and Writing Style in Formal Written Texts, *Text-Interdisciplinary Journal for the Study of Discourse,* 23(3), December 2003.

Coyotl-Morales, R. M.; Villaseñor-Pineda, L.; Montes-y-Gòmez, M.; and Rosso, P. 2006. Authorship attribution using word sequences. In *Proceedings of 11th Ibero-american Congress on Pattern Recognition,* Cancun, Mexico (pp. 844–853).

Harley, B.; and King M. L. 1989. Verb Lexis in the Written Compositions of Young L2 Learners, *Studies in Second Language Acquisition* Volume 11, Issue 4 December, pp. 415-439. https://doi.org/10.1017/S0272263100008421

Hyltenstam, K., 1988. Lexical characteristics of near-native second-language learners of Swedish, *The Journal of Multilingual & Multicultural Development,* 9 (1-2), 67-84.

Kurdi, M. Zakaria. 2016. *Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax*. London: ISTE-Wiley. ISBN-10: 1848218486.

Kurdi, M. Zakaria. 2017a. *Natural Language Processing and Computational Linguistics 2: Semantics, Discourse and Applications*, London: ISTE-Wiley. ISBN: 1848219210.

Kurdi, M. Zakaria. 2017b. Lexical and Syntactic Features Selection for an Adaptive Reading Recommendation System Based on Text Complexity, In *Proceedings of the 2017 International Conference on Information System and Data Mining*, Pages 66-69, Charleston, SC, USA - April 01 - 03, 2017.

Newman, M. L., Groom, C.J., Handelman, L.D. and Pennebaker, J. W. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45: 211-236.

Read, J. 2000. *Assessing vocabulary.* Oxford: Oxford University Press.

Sarawgi, R.; Gajulapalli K.; Choi Y. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *proceedings of the Fifteenth Conference on Computational Natural Language Learning,* Pages 78-86, Portland, Oregon — June 23 - 24.

Shrestha, P.; Bethard, S.; Pedersen T., Rey-Villamizar, N. Farig S., Solorio, T. 2016. Age and Gender Prediction on Health Forum Data. In *proceedings of Language Resources and Evaluation Conference (LREC)*, 23-28 May, Portorož, Slovenia.