Predicting Adolescents' Educational Track from Chat Messages on Dutch Social Media

Lisa Hilte, Walter Daelemans and **Reinhild Vandekerckhove**
CLiPS, University of Antwerp
Prinsstraat 13, 2000 Antwerp, Belgium
{firstname.lastname}@uantwerpen.be

Abstract

We aim to predict Flemish adolescents' educational track based on their Dutch social media writing. We distinguish between the three main types of Belgian secondary education: General (theory-oriented), Vocational (practice-oriented), and Technical Secondary Education (hybrid). The best results are obtained with a Naive Bayes model, i.e. an F-score of 0.68 (std. dev. 0.05) in 10-fold cross-validation experiments on the training data and an F-score of 0.60 on unseen data. Many of the most informative features are character n-grams containing specific occurrences of chatspeak phenomena such as emoticons. While the detection of the most theory- and practice-oriented educational tracks seems to be a relatively easy task, the hybrid Technical level appears to be much harder to capture based on online writing style, as expected.

1 Introduction

While some social variables, such as gender and age, have often been studied in author profiling (see e.g. the overview paper by Reddy et al. (2016)), educational track remains largely unexplored in this respect. The goal of this paper is twofold: we aim to develop a model that accurately predicts adolescents' educational track based on their language use in social media writing, and gain more insight in the linguistic characteristics of youngsters' educational background through inspection of the most informative features for this classification task.

The paper is structured as follows: we start by discussing related research (Section 2). Next, we describe the corpus, as well as the three main types of Belgian secondary education, i.e. the three class labels in the classification experiments (Section 3). Finally, we discuss our methodology (Section 4) and present the results (Section 5).

2 Related Research

Related work on this topic is scarce; only some studies in education profiling can be found, and they examine the impact of tertiary (and not secondary) education, on text genres other than social media writing. Furthermore, Dutch is never the language of interest. Estival et al. (2007), for instance, approached tertiary education profiling as a binary classification task (none versus some tertiary education) for a corpus of English emails. They obtained promising results with an ensemble learner (Bagging algorithm) using character-based, lexical and structural text features while explicitly excluding function words. Pennebaker et al. (2014), however, stressed the importance of function words in a related task: they linked students' writing in college admission essays to their later performance in college. Obtaining higher or lower grades appeared to be associated with the use of certain function words, belonging to either 'categorical' or 'dynamic' writing styles. In previous work on language and social status, Pennebaker (2011) had already pointed out the importance of pronouns: he described a more frequent use of you- and we-words as more typical of high status, as well as a less frequent use of I-words.

When we expand the scope of previous research from profiling studies to other related linguistic fields, we again conclude that this specific topic is underresearched. There are many studies on the characteristics of (youngsters') computer-mediated communication (CMC) (see e.g. Varnhagen et al. (2010), Tagliamonte and Denis (2008) and many more) and even some on the interaction between CMC and education (see e.g. Vandekerckhove and Sandra (2016) for the impact of CMC on school writing). However, the impact of educational track on adolescents' online writing is not addressed. For this specific topic, we can - to our

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 328–334
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

Educational track	Participants	Posts	Tokens
General Secondary Education	596 (43%)	120 839 (28%)	739 831 (29%)
Technical Secondary Education	393 (28%)	197 534 (45%)	1 151 684 (46%)
Vocational Secondary Education	395 (29%)	116 164 (27%)	639 839 (25%)
Total	1 384	434 537	2 531 354

Table 1: Distributions in the corpus.

knowledge - only refer to our previous sociolinguistic work focusing on youngsters with distinct secondary education profiles, in which we have shown that teenagers in practice-oriented tracks tend to deviate more from formal standard writing on social media, by using more typographical chatspeak features (e.g. emoji), more non-standard lexemes (e.g. dialect words) and more non-standard abbreviations (Hilte et al., 2018a,b). While for all examined linguistic features, these differences were very consistent between the two 'poles' of the continuum between theory and practice, i.e. General and Vocational students, the Technical students did not always hold an intermediate position, but their chat messages showed a rather unpredictable linguistic pattern (Hilte et al., 2018a,b). We investigate in this paper whether these sociolinguistic results are confirmed in machine learning experiments.

3 Data Collection

Our corpus consists of Flemish[1] adolescents' private chat messages, written in Dutch on the social media platforms Facebook Messenger and WhatsApp. The data were collected through school visits during which the students were informed about the research, and could voluntarily donate chat messages. We asked for the students' (and for minors, their parents') consent to store and analyze their anonymized texts.

The final corpus contains 434 537 chat messages (2 531 354 tokens) by 1384 authors. All authors are Flemish high school students, aged 13-20, attending one of the three main types of Belgian secondary education: the theory-oriented General Secondary Education (which prepares for higher education), the practice-oriented Vocational Education (which prepares for a specific manual profession) and the hybrid Technical Education, which has both a strong theoretical and practical focus (Flemish Ministry of Education and Training, 2017). An overview of the distributions in the corpus can be found in Table 1.

We note that the Belgian secondary school system is similar to that of several other countries. The distinction between a vocational and an academic training is quite common (e.g. in Denmark, Finland, Croatia, France, Paraguay, China, etc.). The division between three main tracks (offering a more general, technical and vocational program respectively) is made in several countries as well (e.g. Czech Republic, Italy, Turkey, etc.)[2]. Consequently, the present classification task transcends the Belgian context and may be relevant in different countries and cultures, too.

4 Methodology

In this section, we describe the preprocessing of the data and the feature design (resp. Sections 4.1 and 4.2) as well as the experimental setup (Section 4.3).

4.1 Preprocessing

Since we will predict educational track on a participant-level, we must ensure to have sufficient data (and thus a fairly representative sample of online writing) for each participant. For this purpose, we deleted the participants who donated fewer than 50 chat messages. Next, we divided the remaining corpus in a training set (70% of the participants), and a test set (15%). A second test set (15%) was put aside for future experiments. This division was random but stratified, i.e. every subset contained the same proportion of participants per educational track.

4.2 Feature Design

The features used in the classification experiments consist of general textual features and features representing the frequency of typical chatspeak phenomena.

The general features include frequencies for token

[1] I.e. living in Flanders, the Dutch-speaking part of Belgium.

[2] `en.wikipedia.org/wiki/List_of_secondary_education_systems_by_country`

n-grams (uni-, bi- and trigrams) and character n-grams (bi-, tri- and tetragrams). In addition, average token and post length and vocabulary richness (type/token ratio) are taken into account as well. Finally, we use the dictionary-based computational tool LIWC (Pennebaker et al., 2001) in an adaptation for Dutch by Zijlstra et al. (2004) to count word frequencies for semantic and grammatical categories. While counts for individual words are already captured by the token unigrams, these counts per category can allow for broader generalizations for words which are semantically or functionally related. However, we note that the accuracy of this feature might not be optimal, as the social media texts are very noisy (and contain many non-standard elements, e.g. in terms of orthography or lexicon), whereas LIWC is based on standard Dutch word lists.

The set of chatspeak features contains counts for occurrences of several typographic phenomena. It includes the number of character repetitions (e.g. 'suuuuuper nice!!!') and combinations of question and exclamation marks (e.g. 'what?!'). The number of unconventionally capitalized tokens is added as well (alternating, inverse or all caps, e.g. 'AWESOME'). The final typographic features are emoticons and emoji (e.g. :), <3), the rendition of kisses and hugs (e.g. 'xoxoxo'), hashtags for topic indication (e.g. '#addicted') and 'mentions' for addressing a specific person in a group conversation (e.g. '@sarah'). We also add an onomatopoeic variable, i.e. the number of renditions of laughter (e.g. 'hahahahah'). Another typical element of chatspeak are non-standard abbreviations and acronyms (e.g. 'brb' for 'be right back'). The final feature concerns language or register choice per token, in order to explicitly take into account the authors' use of words in a different language or linguistic variety than standard Dutch. We count the number of standard Dutch, English, and non-standard Dutch (e.g. dialect) lexemes. While the other chatspeak features are detected with regular expressions (typographic and onomatopoeic markers) or predefined lists (abbreviations), this lexical feature is extracted using a dictionary-based pipeline approach. For each token, we first checked if it was an actual word (and not e.g. an emoticon). Next, we checked if it occurred in a list of standard Dutch words and named entities. If not, we checked its presence in a standard English word list. Finally, if the token was absent again, it was placed in the 'non-standard Dutch' category. Figure 1 shows a sample of authentic chat messages from the corpus, illustrating the use of several chatspeak features.

#verslaafd	('#addicted')
Neeeeee 😢❤️	('Noooooo')
Haha 😂❤️	
HAHAHAHAHAHAHAHAHAHAHA	

Figure 1: Example messages from the corpus.

For each participant, an individual feature vector was created containing the counts for all of these features. We proceeded with relative counts (to normalize for submission size) by dividing the absolute counts by the author's total number of tokens (e.g. for token unigrams, emoji,) or n-grams (for n-gram frequencies). For initial dimensionality reduction, we applied a frequency cutoff, only taking features into account that are used at least 10 times in the corpus, by at least 5 different participants.

4.3 Experimental Setup

We compared different models to predict Flemish adolescents' educational track based on their social media messages. The classification algorithms we tested were: Support Vector Machines, Naive Bayes (Multinomial, Gaussian and Bernoulli), Decision Trees, Random Forest, and Linear Regression. For all classifiers, we used the Scikit-learn implementation (Pedregosa et al., 2011). For each model, we searched for the optimal parameter settings through a randomized cross-validation search on the training data. We searched for optimal values for classifier-bound parameters (e.g. kernel for SVM), as well as an optimal feature scaler (no scaling, MinMax scaling or binarization) and an optimal percentile for univariate (chi-square based) feature selection, chosen from a continuous distribution. We compared the models' performance in 10-fold cross-validation experiments on the training data.

5 Results

In Section 5.1, we discuss the best model resulting from the 10-fold cross-validation experiments on the training data and compare it to different baseline models. In addition, we inspect the most informative features for the task. In Section 5.2, we discuss additional experiments which provide further insight in the classification problem.

Class levels	Precision	Recall	F-score
General	0.67	0.78	0.72
Technical	0.70	0.54	0.61
Vocational	0.68	0.71	0.70
Avg/total	0.68	0.68	0.68

Table 2: Classification report (in cross-validation).

		Predicted class		
		Gen.	Tech.	Voc.
Actual class	Gen.	153	22	22
	Tech.	49	89	27
	Voc.	25	17	105

Table 3: Confusion matrix (in cross-validation).

5.1 Model Performance and Feature Inspection

The best performing model in CV-setting on the training data is a Multinomial Naive Bayes classifier, with optimized parameters: the value for the smoothing parameter alpha is 0.98, and the model uses the 12.50% best features (according to chi-square tests). The features were binarized. The classification report (Table 2) indicates that the performance is good, with a value of 0.68 for (prevalence-weighted macro-average) precision, recall and F-score (std. dev. 0.05). While precision is very similar for the three educational levels, recall is good for General Education, but slightly worse for the Vocational and much worse for the Technical level. Consequently, the model seems to miss many Technical profiles, confusing them with the other educational tracks. The confusion matrix (Table 3) shows that most (64%) misclassified Technical profiles were incorrectly labeled as the more theory-oriented General track, rather than as the more practice-oriented Vocational track (36%).

As Table 5 summarizes, the model strongly outperforms a probabilistic baseline (0.34) in cross-validation, as well as a simple bag-of-words model (which only uses token unigrams as features) without any parameter tuning, scaling or feature selection (F-score = 0.22). However, when parameter tuning, scaling and feature selection are introduced, the BoW-model obtains almost identical scores in cross-validation: it yields an overall precision, recall and F-score of 0.67 (std. dev. 0.03). There is, however, a difference in how well both models generalize to unseen data. While the first model reaches an average F-score of 0.60 (see Table 4 for the detailed classification report), the BoW-model achieves a lower score of 0.55, and particularly underperforms in the detection of Technical profiles, with an F-score of 0.38 (vs 0.50 for the full model).

Class levels	Precision	Recall	F-score
General	0.64	0.69	0.67
Technical	0.57	0.44	0.50
Vocational	0.58	0.68	0.63
Avg/total	0.60	0.61	0.60

Table 4: Classification report (on unseen data).

In order to better understand the differences and similarities between both models, we compared their feature sets (after feature selection was applied) and inspected the 1000 most informative ones, using information gain as ranking criterion. While we expected that the most informative features for the BoW-model would be lexical and the ones for the full model stylistic, this analysis suggests that in both models, many of the most informative selected features are specific occurrences of chatspeak markers. For the BoW-model, which uses only token unigrams as features, many of the most informative tokens contain one or more chatspeak features (e.g. colloquial register, a spelling manipulation, an emoticon, character repetition, etc.). Some other informative tokens seem to be more content- than style-related, revealing topics such as hobbies, specific locations, friends and school. Strikingly, although the full model contains abstraction of chatspeak phenomena (e.g. total count for emoticons), specific occurrences of these genre markers are still most informative. The 1000 most informative features are all character n-grams: only some reveal topics (e.g. school), but many more indicate the use of chatspeak features, and particularly combinations of emoji/emoticons. Other n-grams indicate the use of English and Arabic words, of colloquial terms, of chatspeak spelling, abbreviations and character repetition. As opposed to the BoW-model's token unigrams, these character n-grams allow the model to capture stylistic features on a sub-token level (e.g. the n-gram 'sss' captures repetition of the letter 's' in different words). We can illustrate a clear advantage by the Arabic word 'wallah' (meaning 'I swear on God's name'), which is often used by our participants with Ara-

	Cross-validation			Unseen data		
Model	Precision	Recall	F-score	Precision	Recall	F-score
Best model	**0.68**	**0.68**	**0.68**	**0.60**	**0.61**	**0.60**
BoW (non-finetuned)	0.15	0.39	0.22	0.15	0.39	0.21
BoW (finetuned)	0.67	0.67	0.67	0.55	0.55	0.55
Stylistic	0.65	0.64	0.64	0.59	0.60	0.59
Prob. baseline	0.34	0.34	0.34	0.34	0.34	0.34

Table 5: Comparison of the different models and baselines.

Class levels	Precision	Recall	F-score
General	0.86	0.80	0.83
Vocational	0.75	0.83	0.79
Avg/total	0.82	0.81	0.81

Table 6: Classification report for binary task (in cross-validation).

Class levels	Precision	Recall	F-score
General	0.82	0.79	0.80
Vocational	0.73	0.77	0.75
Avg/total	0.78	0.78	0.78

Table 7: Classification report for binary task (on unseen data).

bic roots, who spell it in many different ways. Because of these alternative spellings, ‘wallah’ does not appear among the most informative tokens in the BoW-model. However, for the full model, several related character n-grams (e.g. ‘wlh’, ‘wll’) do.

Next, we compared the full model to a stylistic model using only chatspeak features (both abstractions and specific occurrences), and no token or character n-grams. This stylistic model performs slightly worse on both the training set (F-score = 0.64, std. dev. 0.04) and unseen data (F-score = 0.59) (see Table 5). However, inspection of the most informative features in this feature set provides further insight in the education profiling task. Many of the most informative features are again specific occurrences of stylistic phenomena (e.g. specific emoticons, specific lexemes containing letter repetition). Some abstract representations of online writing style characteristics appear among the top-1000 features too (such as the total use of character repetition, of onomatopoeic laughter, acronyms, English words, mentions and hashtags, and emoticons), but much less prominently. These findings suggest that even in a purely stylistic model, abstract representation of certain style features is not informative enough for education profiling, and appears to be less important than the use of these features within specific tokens or contexts.

5.2 Additional Experiments

Additional experiments indicate that the task becomes much easier when the hybrid Technical Education level is not included. Performance for this binary classification task (distinguishing between General and Vocational students only) is much higher (F-score = 0.81 with std. dev. 0.04 in cross-validation, and 0.78 on unseen data; see Tables 6 and 7 for the classification reports), showing that Vocational and General students are not often linguistically confused by the model. Strikingly, in this setting, the purely stylistic model performs similarly on the training data (F-score = 0.81, std. dev. 0.08), and even better on the unseen data (F-score = 0.82) than the full model. This suggests that stylistic differences are more outspoken and consistent between General and Vocational students, and might be sufficient for classification.

Finally, first experiments with separate classifiers for girls and for boys, and for younger versus older teenagers, suggest interesting distinctions (see Table 8). It appears to be easier to correctly predict educational track for girls (F-score = 0.67 with std. dev. 0.07 in cross-validation; and 0.69 on unseen data) than for boys (F-score = 0.60 with std. dev. 0.09 in cross-validation; and 0.66 on unseen data). This suggests that more education-based linguistic variation can be found among girls than among boys. Similarly, better predictions could be made on unseen data for older teenagers, aged 17-20 (F-score = 0.62 in cross-validation, std. dev. 0.07; and 0.63 on unseen data), than for younger

	Cross-validation			**Unseen data**		
Model	Precision	Recall	F-score	Precision	Recall	F-score
Girls	0.67	0.67	0.67	**0.69**	**0.69**	**0.69**
Boys	0.61	0.61	0.60	0.67	0.67	0.66
Younger	**0.69**	**0.69**	**0.69**	0.55	0.55	0.55
Older	0.62	0.62	0.62	0.63	0.63	0.63

Table 8: Comparison of the models for separate groups.

adolescents, aged 13-16 (F-score = 0.69 in cross-validation, std. dev. 0.09; and 0.55 on unseen data). This might be due to the fact that the older teenagers have been together in the same peer networks and class groups for a longer time, and might write more similarly on social media. Furthermore, some of the younger students might actually still change educational track.

6 Conclusion

We conducted classification experiments to predict educational track for Flemish adolescents, based on their social media writing. These first results are promising and indicate that the task is doable. However, although the best model strongly outperforms a probabilistic baseline, its performance is similar to that of a simple BoW-model. This might give the impression that lexical features are still very important; however, inspection of the most informative features revealed that many of the most informative tokens contain stylistic features typical of the informal online genre. The most informative features for the full model suggest that abstraction of these stylistic chatspeak features (or at least, the current implementation) is still of lesser importance than specific occurrences.

While the distinction between General and Vocational high school students appears to be relatively easy to make, the detection of students in the intermediate Technical track is much harder. This could indicate that these students are truly a hybrid class with subsets of students that are simply not that different from their peers in more theory- or more practice-oriented tracks, respectively. In addition, related research shows that these students' online writing is rather unpredictable and does not follow a clear pattern (Hilte et al., 2018a,b).

In future work, we want to experiment with additional algorithms, such as ensemble methods, and with a post-level rather than a participant-level approach (in order to have more data samples at our disposal). We also want to improve the current feature design and particularly the abstract representation of style features, because as van der Goot et al. (2018) write, abstract features may increase generalizability to other corpora (and even genres and languages) in author profiling tasks, compared to lexical models. Finally, we want to further investigate the creation of different classifiers for different subgroups of participants (e.g. boys versus girls).

Finally, we stress that this profiling task is not only relevant in a Belgian context, since the educational tracks serving as class labels correspond to several countries' secondary education programs. Furthermore, the inclusion of stylistic features - i.e. chatspeak phenomena occurring in *any* language - adds to this generalizability. While specific lexemes or specific realizations of chatspeak markers may not always be relevant in other languages or corpora, the abstract stylistic features are more universal on social media. We argue that these models for education profiling, when further improved, could be used in different languages and applications. For instance, the addition of an educational compound can increase existing profiling tools' performance, which can be important in different tasks (e.g. the detection of fake accounts on social media, and many more).

7 Supplementary Materials

Because of the decision of our university's ethical committee, in line with European regulations to ensure the adolescents' privacy, we cannot make the dataset publicly available. The code will be made available.

8 Acknowledgments

We thank Stéphan Tulkens for his advice on the setup and analyses. We are also grateful towards the two anonymous reviewers for their feedback on a previous version of this paper.

References

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272.

Flemish Ministry of Education and Training. 2017. *Statistisch jaarboek van het Vlaams onderwijs. Schooljaar 2015-2016*. Department of Education and Training, Brussels.

Lisa Hilte, Reinhild Vandekerckhove, and Walter Daelemans. 2018a. Adolescents' social background and non-standard writing in online communication. *Dutch Journal of Applied Linguistics*, 7(1):2–25.

Lisa Hilte, Reinhild Vandekerckhove, and Walter Daelemans. 2018b. Social media writing and social class: A correlational analysis of adolescent CMC and social background. *International Journal of Society, Culture & Language*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830.

James W Pennebaker. 2011. *The secret life of pronouns. What our words say about us*. Bloomsbury Press, New York.

James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

T. Taghunadha Reddy, B. Vishnu Vardhan, and P. Vijayapal Reddy. 2016. A survey on authorship profiling techniques. *International Journal of Applied Engineering Research*, 11(5):3092–3102.

Sali A Tagliamonte and Derek Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American speech*, 83(1):3–34.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 383–389.

Reinhild Vandekerckhove and Dominiek Sandra. 2016. De potentiële impact van informele online communicatie op de spellingpraktijk van Vlaamse tieners in schoolcontext. *Tijdschrift voor Taalbeheersing*, 38(3):201–234.

Connie K Varnhagen, G Peggy McFall, Nicole Pugh, Lisa Routledge, Heather Sumida-MacDonald, and Trudy E Kwong. 2010. Lol: New language and spelling in instant messaging. *Reading and writing*, 23(6):719–733.

Hanna Zijlstra, Tanja Van Meerveld, Henriët Van Middendorp, James W Pennebaker, and RD Geenen. 2004. De Nederlandse versie van de 'linguistic inquiry and word count'(LIWC). *Gedrag Gezond*, 32:271–281.