reports

Anonymous ACL submission

1 TEMPLATE

Abstract

For the weekly reports you might not need this, but it's good to get used soon to writing proper abstracts to research papers. Give it a try, and we will provide feedback.

1.1 Introduction

The template is structured along the lines of a research paper, and you can fill each appropriate section with the relevant information. The idea is that you get used to using the standard format adopted in research to report on experiments. At times this might feel a little stretched in the context of homework and the exercises you are asked to complete, but give it a try. Also, don't get too hung up about what should go where: try make decisions, and we will give you feedback. Finally, for spefific assignments you might want to implement some modifications to the template (for example in case you don't have a separate test set, or you might want to have a more general "Experiments" section in case you have to run more than one, and stuff like that). Feel free to do so, as long as you maintain some proper structure which is appropriate for a research paper. Additional questions can be answered in the final section.

1.2 Related work

For the reports you probably don't need this, but in case you want to mention some background or motivate some choices that you made based on the existing literature, this is the place to do it.

1.3 Data

Here you will report on what data you used. How much is it? Where did it come from? Is it annotated? Do you know anything about inter- annotator agreement?

1.4 Method/Approach

Here you will report on method you chose to run your experiments. Also evaluation methods can go here. Any settings you used also can be described here. Do you have a separate dev set? Do you use cross-validation? What features are you using?

1.5 Results

Your final results on test data. You can also include here some results on development, of course, but you should keep them clearly separate from results on test data. Results on development are useful to tune your system, so they better go in Method. But for comparison you might want to report your best dev results also in this section. It can happen that for some assignments you have no separate test set, so this section in case can be merged with the previous one.

1.6 Discussion/Conclusion

What observations can you glean from the results? In the context of the course, you can really use this space not only to discuss the actual results with your own observations, but also to add some reflections on what you had to do, strategies you adopted, what you could do differently, and so on.

1.7 Answers to additional questions

If something doesn't fit in the above, or if there are additional, more general questions in the assignment that are not directly answered in the previous sections, you can use this space for them.

2 Personality prediction from tweets: a language independent approach ~ DRAFT

Abstract

Author profiling is the task of predicting some aspects (e.g., gender, age, personal-

ity) of the author of a given text. In this paper we describe a system that predicts the personality of twitter users. We trained a neural network on the TwiSty corpus and we reached a [TODO ADD] x.xx in RMSE score.

2.1 Introduction

It has been shown that given a text, it is possible to say a lot about its author: its gender, wheter he or she is old or young, if he is depressed, what is his dominant personality trait. This study focuses on this exact last thing: given some text — tweets, in this case — we want to predict the personality of the author.[TODO expand]

Test **?**.

2.2 Related Work

The work of [TODOLIU] shows that neural models can successfully model author personality traits from text. The authors suggest at the end of the paper that "the lack of feature engineering should support language independence". For this work we attempted to replicate the exact same model that [TODO fix][liu et al] built: we wanted to test wether it was possible to use it across languages.

For the baseline: we considered the submissions to the PAN forum [TODO ADD references and footnote] in the last years and we saw that many participants proposed an SVM-based model with a sparse feature representation. Such systems achieved good performances: we decided to use them as a baseline.

2.3 Model

Our system consists of a recurrent and compositional neural-network. A first layer builds embeddings at the character level; the output is used to

2.4 Experiments

2.4.1 Data

For this task we trained and tested our model on the TwiSty corpus [TODO add reference]. Additionally, we worked on the PAN 2015 dataset in order to compare our results with those reported by [TODOLIU]: Table 1 shows a sample of the data.

2.4.2 Pre-processing

For the baseline system we used the default preprocessor included in the scikit-learn tf-idf vectorizer, which lowercases all the words and .tokenizes the text.

2.4.3 Evaluation

2.5 Results

2.5.1 Baseline

2.6 Conclusions and Future Work

Table 1: Sample instances from the PAN 2015 dataset

	author	lang	text	gender	age	ext	sta	agr
0	e5b59ccc-	en	@username @username ay friend, q te fumasteSSS…	F	35-49	0.0	0.2	0.2
1	ed970294-	en	“@username: @username "you can’t have your cak…	M	18-24	0.1	0.2	0.2
2	4b05f4e0-	en	I should probably go to bed considering I have…	F	18-24	0.5	0.0	0.3
3	de7f0515-	en	@username the sameee@username Great!!@u…	M	25-34	0.2	-0.1	0.2
4	a71c93ed-	en	On my very last Nerve!am nothing and I hav…	F	25-34	0.2	0.0	0.0

Table 2: Results (negative MSE and standard deviation, CV-10) for the baseline system on the PAN 2015 dataset using the SVM classifier and unigrams with tf-idf normalization.

	agr	con	ext	opn	sta
en	-0.02 (0.02)	-0.02 (0.02)	-0.02 (0.02)	-0.02 (0.01)	-0.04 (0.04)
es	-0.02 (0.02)	-0.02 (0.03)	-0.02 (0.03)	-0.02 (0.03)	-0.03 (0.03)
it	-0.02 (0.04)	-0.01 (0.01)	-0.02 (0.05)	-0.02 (0.03)	-0.03 (0.04)
nl	-0.03 (0.04)	-0.01 (0.03)	-0.02 (0.03)	-0.01 (0.02)	-0.03 (0.06)