Instructions for ACL-2018 Proceedings

Abstract

This document contains the instructions for preparing a camera-ready manuscript for the proceedings of ACL-2018. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used for both papers submitted for review and for final versions of accepted papers. Authors are asked to conform to all the directions reported in this document.

1. Introduction

Natural language processing is a complex task, one project that has been attempted many times is author profiling [1,2,3]. Author Profiling is a method of analysing a set of texts in order to identify various characteristics of the author, in this particular project, the research question that will be addressed is; ‘Can Age and Gender be inferred from Social Media Posts?’, the two key characteristics that will be determined within the author profiling are birth year and gender.

Author profiling is of key importance in forensic linguistics, marketing and data protection [1,9]. The use of NLP in order to identify a person’s age and gender can be used to prevent child groomers in online chat rooms, in addition to this, new data protection laws introduce a barrier to the information marketers have on the average consumer, however if a user’s age or gender can be inferred from their social media posts, advertisers could use this for targeted marketing. People who are close in age have similar interests, and as a result have similar age-related life events, using this logic, the tweets of a person aged 20 should be closer to someone within 5 years of them as opposed to a person 10 years away from them. Similarly with gender, language features, which are words or phrases used in texts such as metaphors, alliteration and oxymorons are gender-preferential, as opposed to being exclusive to one sex [4]. An example of this being the responses of men and women, both men and women provide minimal responses when listening to a speaker however the context, timing and frequency of the responses vary with gender [5, 6]. However, it is also noted that people of the same gender use different gender preferential features than other members of the same gender, an example of this being members of the same gender such as men speaking in long utterance lengths, whereas other men may use a high number of insults and aggressive utterences, this highlights the challenge of differentiating between genders. Language is always changing, evolving and adapting to the needs of it’s users, and differentiating people’s age based on social media posts can be a challenge due to a person’s age constantly changing along with a persons speech evolving as time goes on. As opposed to identifying a person’s age, as this is relative to the time an age is queried, the birth-year of each user will be predicted instead. In addition to this, a user’s gender will be also predicted, which will be binary result; either male or female in order to answer the question; Can Age and Gender be predicted using text obtained from Social Media posts?

1. Related Work

In [1], author profiling was achieved by studying sociolect aspect, which is how language is shared by people. A blog corpus was used with N-grams of POS tags in documents, punctuation symbols and number of links with an SVM was used to determine unknown author’s age and gender that achieved a 64.30% accuracy with age and a 64.73% accuracy for gender. In [1], the only metric that was used for classification evaluation was accuracy, and the data was unevenly distributed with a significantly larger proportion of the corpus coming from ages 25 and above. In [2] a corpus of chat texts was collected from a Belgian social networking site called Netlog. The text was tokenized and the feature set was limited by token and character features only, the feature sets were built by selecting features with the highest χ2 value and an SVM classifier was used to classify age which achieved an accuracy of 71.3%, and when classifying age and gender an accuracy of 66.3% was achived on their whole dataset, it was found that token features outperformed character features in each age classification task. This paper achieved a highest accuracy score of 88.8%, precision score of 91.5%, recall score of 92.9% and best F-score of 91.7% for the adult class of a smaller subset of data. This paper tested the models on data taken from the same site the model was trained on and although the accuracies achieved were 88% and above the data used was non-standard language specific to the Netlog site written by users in Belgium with access to Netlog specific emojis.

A study was conducted [3] where particapants analysed a text that was 50 words long regarding a gender neutral topic, each text had small changes made to them in order to create a male version and a female version of the text and each particapant had to identify 3 different texts as either male or female. It was found that participants could correctly discriminate gender of users, on the basis of linguistic-style when other cues to gender were absent, highlighting the importance of lingustic-style when identifying gender.

1. Data

Within this project, tweets from a user’s twitter page was gathered and seperated by gender, and birth year. All of the data that was used in this project originated from PAN [7]. In order to obtain a representitive sample, it was ensured the dataset that was selected contained tweets that were gathered from USA, Australia, UK and Canada. Obtaining a large dataset that contained verified birth-year and gender tags for a series of text proved very challenging, however the data sourced in this project [7], was deemed suitable for the question proposed, as it contains 48,335 user profiles, with verified birth-year and gender tags, and the number of associated tweets contained with each profile are of significant length. However, it is noted that this study is limited by the countries and areas that were selected to gather data from, in addition to the fact that this project has a focus on only tweets written in English. This is a limitation due to the fact that the composure of tweets by Men and Women of different ages may significantly vary when writing in different languages and further inferences could be gained from this information which could significantly improve the models used. Due to many people falsifying their age on social media [CITE] the majority of data that was collected from twitter was from users that were verified on twitter [CITE], all of the data was from celebrities, this ensured the data used to train the models was an accurate representation of text written by the associated age and gender. It is also noted that this data is from celebrities with fame ranging from rising, star and superstar, although the fame level is varied, this dataset is limited by only the tweets that celebrities post, and not how the average non-celebrity may tweet. Another limitation is the the fact that this textual data is only from users that have a twitter account, the models used have not been exposed to text outside of twitter, in addition to this, each tweet can have a maximum of 180 chars, which will influence how people write posts on twitter, and hence influence the classifiers used.

1. Exploratory Data Analysis

Here, exploratory data analysis has been carried out in order to discover and identify patterns, spot anomalies, check the class distribution and check any assumptions made by producing summary statistics and simple, easily intepretable graphs.

Hist of Age

BarPlot of Age

Bar Plot of Gender

tween each line is Multiple, 1.05 in, with 4.5 pt below each line. After pasting a new ruler onto a page, highlight the ruler text and press **F9** to update the numbers. **Align** the text box **Middle**, and **Center**, and then **Arrange** Behind Text.**Reviewers:** Note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (*e.g.*, this paragraph ends at mark 141.5).h 4.5 pt below each line. After pasting a new ruler onto a page, highlight the ruler text and press **F9** to update the numbers. **Align** the text box **Middle**, and **Center**, and then **Arrange** Behind Text.**Reviewers:** Note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (*e.g.*, this paragraph ends at mark 141.5).h 4.5 pt below each line. After pasting a new ruler onto a page, highlight the ruler text and press **F9** to update the numbers. **Align** the text box **Middle**, and **Center**, and then **Arrange** Behind Text.**Reviewers:** Note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (*e.g.*, this paragraph ends at mark 141.5).

1. Ethics

Teenagers have now started to use social media as their main means of communication [11]. Moreover a recent study of adolescents and mobile phones [12] ascertained 70% of 11-12 year olds in the UK now own a mobile phone, this figure is rising to 90% by age 14 [12]. As a result, child grooming has increased and the number of cases every year is on the rise. It is noted that this project could be used by people with bad intentions, by identifying ages/genders of users that wish to remain anonymous. For ecample some users may be young or do not wish to be targeted by certain adverts and by accurately predicting a user’s age and gender this information would not longer be anonymous, with this in mind, this classifier and any associated code will be distributed with caution.

1. Methodology

For the production of the electronic manuscript, you must use Adobe's Portable Document Format (PDF). This format can be generated from postscript files: on Unix systems, you can use ps2pdf for this purpose; on Windows, you can use Adobe's Distiller, an online service such as <http://go4convert.com/>, or if you have cygwin installed, you can use dvipdf or ps2pdf.

Please make sure that your PDF file includes all the necessary formatting, hyperlinks, and fonts (especially tree diagrams, symbols, and fonts with Asian characters). When you print or create the PDF file, there is usually an option in your printer setup to include none, all, or just non-standard fonts. Please make sure that you select the option of including ALL the fonts. **Before sending it, test your PDF by printing it from a computer different from the one where it was created.** Moreover, some word processors may generate very large PDF files, where each page is rendered as an image. Such images may reproduce poorly. In this case, try alternative ways to obtain the PDF. One way on some systems is to install a driver for a postscript printer, send your document to the printer specifying “Output to a file”, then convert the file to PDF.

It is of utmost importance to specify the **A4 format** (21 cm x 29.7 cm) when formatting the paper. Print-outs of the PDF file on A4 paper should be identical to the hardcopy version. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs above as soon as possible.

1. Results and Findings
2. Conclusion and Future Work

For reasons of uniformity, Adobe's **Times Roman** font should be used. If Times Roman is not available, you may use the Times New Roman font, which is often provided by default and only slightly different.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

References

[1] Santosh, K., Bansal, R., Shekhar, M. and Varma, V., 2013. Author profiling: Predicting age and gender from blogs. Notebook for PAN at CLEF, 2013.

[2] Peersman, C., Daelemans, W. and Van Vaerenbergh, L., 2011, October. Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents (pp. 37-44). ACM.

[3] Thomson, R. and Murachver, T., 2001. Predicting gender from electronic discourse. British Journal of Social Psychology, 40(2), pp.193-208.

[4] Fitzpatrick, Mary & Mulac, Anthony & Dindia, Kathryn. (1995). Gender-Preferential Language Use in Spouse and Stranger Interaction. Journal of Language and Social Psychology - J LANG SOC PSYCHOL. 14. 18-39. 10.1177/0261927X95141002.

[5] Hannah, A., & Murachver, T. (1999). Gender and Conversational Style as Predictors of Conversational Behavior. Journal of Language and Social Psychology, 18(2), 153–174. https://doi.org/10.1177/0261927X99018002002

[6] Zimmerman, D. H., and West, C., (1975) Sex roles, interruptions and silences in conversation Language and sex: Difference and dominance. pp: 105- 129. Stanford, CA: Stanford University Press.

[7] Pan.webis.de. (2019). PAN @ CLEF 2019 - Celebrity Profiling. [online] Available at: https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html [Accessed 9 Apr. 2019].

[8] Álvarez-Carmona, Miguel Á., et al. "Evaluating topic-based representations for author profiling in social media". Ibero-American Conference on Artificial Intelligence. Springer, Cham, 2016. p. 151-162.

[9] Santosh, Kosgi & Bansal, Romil & Shekhar, Mihir & Varma, Vasudeva. (2013). Author Profiling : Predicting Age and Gender from Blogs Notebook for PAN at CLEF 2013.

[10] Cano, A.E., Fernandez, M. and Alani, H., 2014, November. Detecting child grooming behaviour patterns on social media. In International conference on social informatics (pp. 412-427). Springer, Cham.

[11] Australian Institute of Criminology (AIC). Children’s use of mobile phones. Technical report, GSMA, NTT DOCOMO (2013)

[12] SCAMP Project. Study of cognition, adolescents and mobile phones (scamp). Technical report, Imperial College London (2014)