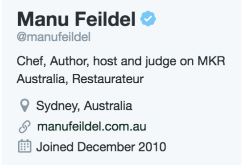
Detecting Social Roles in Twitter

Sunghwan Mac Kim, Stephen Wan and **Cécile Paris**
Data61, CSIRO, Sydney, Australia
{Mac.Kim, Stephen.Wan, Cecile.Paris}@csiro.au

Abstract

For social media analysts or social scientists interested in better understanding an audience or demographic cohort, being able to group social media content by demographic characteristics is a useful mechanism to organise data. Social roles are one particular demographic characteristic, which includes work, recreational, community and familial roles. In our work, we look at the task of detecting social roles from English Twitter profiles. We create a new annotated dataset for this task. The dataset includes approximately 1,000 Twitter profiles annotated with social roles. We also describe a machine learning approach for detecting social roles from Twitter profiles, which can act as a strong baseline for this dataset. Finally, we release a set of word clusters obtained in an unsupervised manner from Twitter profiles. These clusters may be useful for other natural language processing tasks in social media.

1 Introduction



Manu Feildel
@manufeildel

Chef, Author, host and judge on MKR Australia, Restaurateur

Sydney, Australia
manufeildel.com.au
Joined December 2010

Figure 1: An example of a Twitter profile.

Social media platforms such as Twitter have become an important communication medium in society. As such, social scientists and media analysts are increasingly turning to social media as a cheap and large-volume source of real-time data, supplementing "traditional" data sources such as interviews and questionnaires. For these fields, being able to examine demographic factors can be a key part of analyses. However, demographic characteristics are not always available on social media data. Consequently, there has been a growing body of work investigating methods to estimate a variety of demographic characteristics from social media data, such as gender and age on Twitter and Facebook (Mislove et al., 2011; Sap et al., 2014) and YouTube (Filippova, 2012). In this work we focus on estimating social roles, an under-explored area.

In social psychology literature, Augoustinos et al. (2014) provide an overview of schemata for social roles, which includes achieved roles based on the choices of the individual (e.g., writer or artist) and ascribed roles based on the inherent traits of an individual (e.g., teenager or schoolchild). Social roles can represent a variety of categories including gender roles, family roles, occupations, and hobbyist roles. Beller et al. (2014) have explored a set of social roles (e.g., occupation-related and family-related social roles) extracted from the tweets. They used a pragmatic definition for social roles: namely, the word following the simple self-identification pattern "I am a/an ". In contrast, our manually annotated dataset covers a wide range of social roles without using this fixed pattern, since it is not necessarily mentioned before the social roles.

On Twitter, users often list their social roles in their profiles. Figure 1, for example, shows the Twitter profile of a well-known Australian chef, Manu Feildel (@manufeildel). His profile provides infor-

Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, pages 34–40,
Austin, TX, November 1, 2016. ©2016 Association for Computational Linguistics

mation about his social roles beyond simply listing occupations. We can see that he has both a profession, *Chef*, as well as a community role, *Judge* on My Kitchen Rules (MKR), which is an Australian cooking show.

The ability to break down social media insights based on social roles is potentially a powerful tool for social media analysts and social scientists alike. For social media analysts, it provides the opportunity to identify whether they reach their target audience and to understand how subsets of their target audience (segmented by social role) react to various issues. For example, a marketing analyst may want to know what online discussions are due to parents versus other social roles.

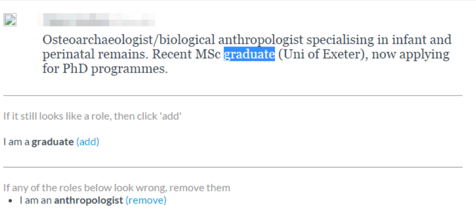
Our aim in this paper is to provide a rich collection of English Twitter profiles for the social role identification task. The dataset includes a approximately 1,000 Twitter profiles, randomly selected, which we annotated with social roles. Additionally, we release unsupervised Twitter word clusters that will be useful for other natural language processing (NLP) tasks in social media.[1] Finally, we investigate social role tagging as a machine learning problem. A machine learning framework is described for detecting social roles in Twitter profiles.

Our contributions are threefold:

- We introduce a new annotated dataset for identifying social roles in Twitter.
- We release a set of Twitter word clusters with respect to social roles.
- We propose a machine learning model as a strong baseline for the task of identifying social roles from Twitter profiles.

2 Crowdsourcing Annotated Data

Twitter user profiles often list a range of interests that they associate with, and these can vary from occupations to hobbies (Beller et al., 2014; Sloan et al., 2015). The aim of our annotation task was to manually identify social role-related words in English Twitter profile descriptions. A social role is defined as a single word that could be extracted from the description. These can include terms such as *engineer*, *mother*, and *fan*. For instance, we obtain *Musician* and *Youtuber* as social roles from "Australian Musician and Youtuber who loves purple!".[2]



[redacted]

Osteoarchaeologist/biological anthropologist specialising in infant and perinatal remains. Recent MSc graduate (Uni of Exeter), now applying for PhD programmes.

If it still looks like a role, then click 'add'

I am a **graduate** (add)

If any of the roles below look wrong, remove them

- I am an **anthropologist** (remove)

Figure 2: The Crowdflower annotation interface.

To study social roles in Twitter profiles, we compiled a dataset of approximately 1,000 randomly selected English Twitter profiles which were annotated with social roles. These samples were drawn from a large number of Twitter profiles crawled by a social network-based method (Dennett et al., 2016). Such a dataset provides a useful collection of profiles for researchers to study social media and to build machine learning models.

Annotations were acquired using the crowdsourcing platform Crowdflower.[3], which we now outline.

2.1 Crowdflower Annotation Guidelines

We asked Crowdflower annotators to identify social roles in the Twitter profiles presented to them, using the following definition: "Social roles are words or phrases that could be pulled out from the profile and inserted into the sentence *I am a/an* ...". Note that the profile does not necessarily need to contain the phrase "I am a/an" before the social role, as described in Section 1.

The annotation interface is presented in Figure 2. The annotator is asked to select spans of text. Once a span of text is selected, the interface copies this text into a temporary list of candidate roles. The annotator can confirm that the span of text should be kept as a role (by clicking the 'add' link which moves the text span to a second list representing the "final candidates"). It is also possible to remove a candidate role from the list of final candidates (by clicking 'remove'). Profiles were allowed to have more than one social role.

Annotators were asked to keep candidate roles as short as possible as in the following instruction: *if*

[1] Our dataset and word clusters are publicly available at `https://data.csiro.au`.

[2] This is a real example.

[3] crowdflower.com

Number of annotated profiles	983
Average description length	13.02 words
Longest description length	74 words
Shortest description length	1 word
Number of unique roles	488

Table 1: Descriptive statistics for the annotated data.

the Twitter profile contains "Bieber fan", just mark the word "fan".[4] Finally, we instructed annotators to only mark roles that refer to the owner of the Twitter profile. For example, annotators were asked not to mark *wife* as a role in: *I love my wife*. Our Crowdflower task was configured to present five annotation jobs in one web page. After each set of five jobs, the annotator could proceed to the next page.

2.2 Crowdflower Parameters

To acquire annotations as quickly as possible, we used the *highest speed* setting in Crowdflower and did not place additional constraints on the annotator selection, such as language, quality and geographic region. The task took approximately 1 week. We offered 15 cents AUD per page. To control annotation quality, we utilised the Crowdflower facility to include test cases called *test validators*, using 50 test cases to evaluate the annotators. We required a minimum accuracy of 70% on test validators.

2.3 Summary of Annotation Process

At the completion of the annotation procedure, Crowdflower reported the following summary statistics that provide insights on the quality of the annotations. The majority of the judgements were sourced from annotators deemed to be *trusted* (i.e., reliable annotators) (4750/4936). Crowdflower reported an inter-annotator agreement of 91.59%. Table 1 presents some descriptive statistics for our annotated dataset. We observe that our Twitter profile dataset contains 488 unique roles.

In Table 2, we present the top 10 ranked social roles. As can be seen, our extracted social roles include terms such as *student* and *fan*, highlighting that social roles in Twitter profiles include a diverse range of personal attributes. In Table 3, we see that more than half (56.2%) of the descriptions do not contain any role, and approximately 22.7% contain one role. The remaining descriptions (21.1%) contain more than one social role.

Social role	Frequency
student	25
fan	24
girl	16
writer	14
teacher	13
geek	12
author	11
artist	10
directioner	9
designer	8

Table 2: Top 10 ranked social roles in Twitter profiles.

Number of roles	Frequency (%)
0	552 (56.2)
1	213 (22.7)
2	101 (10.3)
3	45 (4.6)
4	31 (3.2)
5	23 (2.3)
6	8 (0.8)
7	2 (0.2)
8	6 (0.6)
9	2 (0.2)

Table 3: Frequencies of number of roles that are used to annotate one Twitter profile in our dataset.

3 Word Clusters

We can easily access a large-scale unlabelled dataset using the Twitter API, supplementing our dataset, to apply unsupervised machine learning methods to help in social role tagging. Previous work showed that word clusters derived from an unlabelled dataset can improve the performance of many NLP applications (Koo et al., 2008; Turian et al., 2010; Spitkovsky et al., 2011; Kong et al., 2014). This finding motivates us to use a similar approach to improve tagging performance for Twitter profiles.

Two clustering techniques are employed to generate the cluster features: Brown clustering (Brown et al., 1992) and K-means clustering (MacQueen, 1967). The Brown clustering algorithm induces a hierarchy of words from an unannotated corpus, and it allows us to directly map words to clusters. Word embeddings induced from a neural network are often useful representations of the meaning of words, encoded as distributional vectors. Unlike Brown clustering, word embeddings do not have any form of clusters by default. K-means clustering is thus used on the resulting word vectors. Each word is mapped to the unique cluster ID to which it was assigned, and these cluster identifiers were used as features.

[4] While this decision could give us a coarse-grain granularity of social roles, it was an application-specific requirement from a visualisation point of view to minimise roles.

Bit string	Words related to social role
010110111100	**writer**, nwriter, scribbler, writter, glutton
01011010111110	**teacher**, tutor, preacher, homeschooler, nbct, hod, dutchman, nqt, tchr
0101101111110	**musician**, philologist, orchestrator, memoirist, dramatist, violist, crooner, flautist, filmaker, humourist, dramaturg, harpist, flutist, trumpeter, improvisor, trombonist, musicologist, organist, puppeteer, laureate, poetess, hypnotist, audiobook, comedienne, saxophonist, cellist, scriptwriter, narrator, muso, essayist, improviser, satirist, thespian, ghostwriter, arranger, humorist, violinist, magician, lyricist, playwright, pianist, screenwriter, novelist, performer, philosopher, composer, comedian, filmmaker, poet

Table 4: Examples of Brown clusters with respect to social roles: *writer*, *teacher* and *musician*.

Cluster	Words related to social role
937	**writer**, freelance, interviewer, documentarian, erstwhile, dramaturg, biographer, reviewer, bookseller, essayist, unpublished, critic, author, aspiring, filmmaker, dramatist, playwright, laureate, humorist, screenwriter, storyteller, ghostwriter, copywriter, scriptwriter, proofreader, copyeditor, poet, memoirist, satirist, podcaster, novelist, screenplay, poetess
642	**teacher**, learner, superintendent, pyp, lifelong, flipped, preparatory, cue, yearbook, preschool, intermediate, nwp, school, primary, grades, prek, distinguished, prep, dojo, isd, hpe, ib, esl, substitute, librarian, nbct, efl, headteacher, mfl, hod, elem, principal, sped, graders, nqt, eal, tchr, secondary, tdsb, kindergarten, edd, instructional, elementary, keystone, grade, exemplary, classroom, pdhpe
384	**musician**, songwriter, singer, troubadour, arranger, composer, drummer, session, orchestrator, saxophonist, keyboardist, percussionist, guitarist, soloist, instrumentalist, jingle, trombonist, vocal, backing, virtuoso, bassist, vocalist, pianist, frontman

Table 5: Examples of word2vec clusters with respect to social roles: *writer*, *teacher* and *musician*.

We used 6 million Twitter profiles that were automatically collected by crawling a social network starting from a seed set of Twitter accounts (Dennett et al., 2016) to derive the Brown clusters and word embeddings for this domain. For both methods, the text of each profile description was normalised to be in lowercase and tokenised using whitespace and punctuation as delimiters.

To obtain the Brown clusters, we use a publicly available toolkit, *wcluster*[5] to generate 1,000 clusters with the minimum occurrence of 40, yielding 47,167 word types. The clusters are hierarchically structured as a binary tree. Each word belongs to one cluster, and the path from the word to the root of the tree can be represented as a bit string. These can be truncated to refer to clusters higher up in the tree.

To obtain word embeddings, we used the skip-gram model as implemented in *word2vec*[6], a neural network toolkit introduced by (Mikolov et al., 2013), to generate a 300-dimension word vector based on a 10-word context window size. We then used K-means clustering on the resulting 47,167 word vectors (k=1,000). Each word was mapped to the unique cluster ID to which it was assigned.

Tables 4 and 5 show some examples of Brown clusters and word2vec clusters respectively, for three social roles: *writer*, *teacher* and *musician*. We note that similar types of social roles are grouped into the same clusters in both methods. For instance, *orchestrator* and *saxophonist* are in the same cluster containing *musician*. Both clusters are able to capture the similarities of abbreviations of importance to social roles, for example, *tchr → teacher*, *nbct → National Board Certified Teachers*, *hpe → Health and Physical Education*.

4 Identifying Social Roles

4.1 Social Role Tagger

This section describes a tagger we developed for the task of identifying social roles given Twitter profiles. Here, we treat social role tagging as a sequence labelling task. We use the MALLET toolkit (McCallum, 2002) implementation of Conditional Random Fields (CRFs) (Lafferty et al., 2001) to automatically identify social roles in Twitter profiles as our machine learning framework. More specifically, we employ a first-order linear chain CRF, in which the preceding word (and its features) is incorporated as context in the labelling task. In this task, each word is tagged with one of two labels: social roles are tagged with *R* (for "role"), whereas the other words are tagged by *O* (for "other").

The social role tagger uses two categories of features: (i) basic lexical features and (ii) word cluster features. The first category captures lexical cues that may be indicative of a social role. These features include morphological, syntactic, orthographic and regular expression-based features (McCallum and Li, 2003; Finkel et al., 2008). The second captures semantic similarities, as illustrated in Tables 4 and 5 (Section 3). To use Brown clusters in CRFs, we use eight bit string representations of different lengths to create features representing the ancestor clusters of the word. For word2vec clusters, the cluster identifiers are used as features in CRFs. If a word is

[5] `https://github.com/percyliang/brown-cluster`

[6] `https://code.google.com/p/word2vec/`

Model	Feature	Precision	Recall	F1
KWS		0.659	0.759	0.690
CRFs	Basic	0.830	0.648	0.725
	+ Brown	0.859	0.708	0.774
	+ W2V	0.837	0.660	0.736
	+ (Brown+W2V)	0.863	0.712	**0.779**

Table 6: 10-fold cross-validation macro-average results on the annotated dataset. (Brown: Brown cluster features, W2V: Word2vec cluster features).

not associated with any clustering, its corresponding cluster features are set to null in the feature vector for that word.

4.2 Evaluation

We evaluate our tagger on the annotated Twitter dataset using precision, recall and F1-score. We use 10-fold cross-validation and report macro-averages. Significance tests are performed using the Wilcoxon signed-rank test (Wilcoxon, 1945). We compare the CRF-based tagger against a keyword spotting (KWS) method. This baseline uses social roles labelled in the training data to provide keywords to spot for in the test profiles without considering local context. On average, over the 10-fold cross-validation, 54% of the social roles in the test set are seen in the training set. This indicates that the KWS baseline has potential out-of-vocabulary (OOV) problems for unseen social roles.

To reduce overfitting in the CRF, we employ a zero mean Gaussian prior regulariser with one standard deviation. To find the optimal feature weights, we use the limited-memory BFGS (L-BFGS) (Liu and Nocedal, 1989) algorithm, minimising the regularised negative log-likelihood. All CRFs are trained using 500 iterations of L-BFGS with the Gaussian prior variance of 1 and no frequency cutoff for features, inducing approximately 97,300 features. We follow standard approaches in using the forward-backward algorithm for exact inference in CRFs.

Table 6 shows the evaluation results of 10-fold cross-validation for the KWS method and the CRF tagger. With respect to the different feature sets, we find that the combination of the word cluster features obtained by the two methods outperform the basic features in terms of F1 (77.9 vs. 72.5 respectively), in general providing a statistically significant improvement of approximately 5% (p<0.01).

The improvement obtained with word cluster features lends support to the intuition that capturing similarity in vocabulary within the feature space helps with tagging accuracy. Word cluster models provide a means to compare words based on semantic similarity, helping with cases where lexical items in the test set are not found in the training set (e.g., linguist, evangelist, teamster). In addition, the cluster features allow CRFs to detect informal and abbreviated words as social roles. Our tagger identifies both *teacher* and *tchr* as social roles from the two sentences: “I am a school teacher” and “I am a school tchr”. This is particularly useful in social media because of the language variation in vocabulary that is typically found.

In this experiment, we show that social role tagging is possible with a reasonable level of performance (F1 77.9), significantly outperforming the KWS baseline (F1 69.0). This result indicates the need for a method that captures the context surrounding word usage. This allows language patterns to be learned from data that disambiguate word sense and prevents spurious detection of social roles from the data. This is evidenced by the lower precision and F1-score for the KWS baseline, which over-generates candidates for social roles.

5 Conclusion and Future Work

In this work, we constructed a new manually annotated English Twitter profile dataset for social role identification task. In addition, we induced Twitter word clusters from a large unannotated corpus with respect to social roles. We make these resources publicly available in the hope that they will be useful in research on social media. Finally, we developed a social role tagger using CRFs, and this can serve as a strong baseline in this task. In future work, we will look into being able to identify multi-word social roles to obtain a finer-grained categorisation (e.g., “chemical engineer” vs. “software engineer”).

Acknowledgments

The authors are grateful to anonymous SocialNLP@EMNLP reviewers. We would also like to thank David Milne, who ran the data collection and annotation procedure, and Bo Yan, who collected the additional data for the clustering methods.

References

Martha Augoustinos, Iain Walker, and Ngaire Donaghue. 2014. *Social cognition: an integrated introduction*. SAGE London, third edition.

Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. 2014. I'm a belieber: Social roles via self-identification and conceptual attributes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 181–186, Baltimore, Maryland, June. Association for Computational Linguistics.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, December.

Amanda Dennett, Surya Nepal, Cecile Paris, and Bella Robinson. 2016. Tweetripple: Understanding your twitter audience and the impact of your tweets. In *Proceedings of the 2nd IEEE International Conference on Collaboration and Internet Computing*, Pittsburgh, PA, USA, November. IEEE.

Katja Filippova. 2012. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1478–1488, Jeju Island, Korea, July. Association for Computational Linguistics.

Jenny Rose Finkel, Alex Kleeman, and Christopher D. Manning. 2008. Efficient, feature-based, Conditional Random Field parsing. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 959–967, Columbus, Ohio, June. Association for Computational Linguistics.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1001–1012, Doha, Qatar, October. Association for Computational Linguistics.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 595–603, Columbus, Ohio, June. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(3):503–528, Dec.

J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, California. University of California Press.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 188–191, Edmonton, Canada. Association for Computational Linguistics.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June. Association for Computational Linguistics.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 554–557. The AAAI Press.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1151, Doha, Qatar, October. Association for Computational Linguistics.

Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. 2015. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PLoS ONE*, 10(3):e0115545, 03.

Valentin I. Spitkovsky, Hiyan Alshawi, Angel X. Chang, and Daniel Jurafsky. 2011. Unsupervised dependency parsing without gold part-of-speech tags. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1281–1290, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.