Temporal Orientation of Tweets for Predicting Income of Users

Mohammed Hasanuzzaman[1], **Sabyasachi Kamila**[2], **Mandeep Kaur**[2],
Sriparna Saha[2] **and Asif Ekbal**[2]

[1]ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
[2]Department of Computer Science and Engineering,
Indian Institute of Technology Patna, India
hasanuzzaman.im@gmail.com

Abstract

Automatically estimating a user's socio-economic profile from their language use in social media can significantly help social science research and various downstream applications ranging from business to politics. The current paper presents the first study where user cognitive structure is used to build a predictive model of income. In particular, we first develop a classifier using a weakly supervised learning framework to automatically time-tag tweets as past, present, or future. We quantify a user's overall temporal orientation based on their distribution of tweets, and use it to build a predictive model of income. Our analysis uncovers a correlation between future temporal orientation and income. Finally, we measure the predictive power of future temporal orientation on income by performing regression.

1 Introduction

User-generated content in social media such as Twitter has enabled the study of author profiling on an unprecedented scale. Author profiling in social media aims at inferring various attributes of the user from the text that they have written. Most of the prior studies in this field have focused on age, gender prediction (Marquardt et al., 2014; Sap et al., 2014), psychological well-being (Dodds et al., 2011; Choudhury et al., 2013), and a host of other behavioural, psychological and medical phenomena (Kosinski et al., 2013). However, there has been a lack of work looking at the socio-economic characteristics of Twitter users. In this paper, we focus on automatic estimation of Twitter users' income from their Twitter language. An income predictor of social media users can be useful for both social science research and a range of downstream applications in banking, marketing, and politics.

Previous social science studies on income demonstrate that income of people is correlated with various factors such as demographic feature (the congressional district in which the respondent lived), educational categories, sex, age, age squared, gender, race categories, marital status categories, and height (Kahneman and Deaton, 2010). Other studies reveal that psychological traits related to extroversion (e.g. larger social networks) and conscientiousness (e.g. orderliness) have a positive correlation with income, while neurotic traits (e.g. anger, anxiety) are anti-correlated (Roberts et al., 2007). Human temporal orientation refers to individual differences in the relative emphasis one places on the past, present, or future (Zimbardo and Boyd, 2015). Past studies have established consistent links between temporal orientation and most of the above-mentioned income predictor factors such as age, sex, gender, education, and psychological traits (Webley and Nyhus, 2006; Adams and Nettle, 2009; Schwartz et al., 2013; Zimbardo and Boyd, 2015). Accordingly, this begs the question as to whether there is any link between an individual's temporal orientation and their income level. Traditionally, temporal orientation has been assessed by self-report questionnaires. In this paper, we assess temporal orientation based on language use in Twitter. Our method uses a tweet-level classifier of past, present, and future, grouped over users to create user-level assessments.

Our learning framework uses convolutional neural networks (CNNs) (Goodfellow et al., 2016) to infer tweet vector representations, and considers them as the feature to develop a classification model that can automatically detect the time orientation (oriented towards *past, present*, and *future*) of tweets. The framework leverages weak supervision signals provided by a list of manu-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 659–665
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2104

ally selected eighty (80) high-precision seed terms (and automatically extracted similar terms) representing past, present, and future to train the CNN. For example, tweets exclusively containing *past* (resp. *present* and *future*) seed terms were marked with weak labels *past* (resp. *present* and *future*). We used the tweet-level temporal classifier to automatically classify a large dataset consisting of ≈10 million tweets from 5,191 users mapped to their income, using fine-grained user occupation as a proxy. Finally, we tested whether individual differences in past, present, and future orientation are related to income. In particular, we frame the income prediction task as regression using linear as well as non-linear learning algorithms where temporal orientation served as predictive features. To the best of our knowledge, this represents the first work to study a temporal orientation-based income prediction using Twitter language.

In summary the proposed approach is different from the previous works (Schwartz et al., 2015; Preoţiuc-Pietro et al., 2015; Park et al., 2017) in several ways. Unlike Schwartz et al. (2015), we used a weakly supervised approach. The generation of training data is semi-automatic in our case. Rather than manually identifying features, tweet vectors are fed to a CNN classifier. Furthermore while Schwartz et al. (2015) studied temporal orientation of facebook data in order to predict different human correlates like conscientiousness, age, and gender, our current work focuses on predicting the income of a user using temporal orientation of their tweets. In Preoţiuc-Pietro et al. (2015), the authors predict user income based on different demographic and psychological features of users. However, the process of extracting these features is computationally complex. The current study is therefore, the first of its kind to explore the use of temporal orientation of user-tweets to predict income.

2 Related Work

Existing message-/sentence-level temporal classification methods generally fall into two categories: (1) rule-based methods, and (2) supervised machine-learning methods. Rule-based methods mainly rely on manually designed classification rules for each temporal class (Nie et al., 2015). Despite their effectiveness, this kind of method requires substantial efforts in rule design. Most research on machine learning-based sentence temporal classification has revolved around feature engineering for better classification performance. Different kinds of features have been explored such as bag-of-words, time expressions, part-of-speech tags, and temporal class-specific lexicons (Schwartz et al., 2015). Temporal class specific lexicon creation and feature engineering also cost a lot of human efforts. In addition, creation of a large-scale training data set for supervised machine-learning approaches is also very laborious.

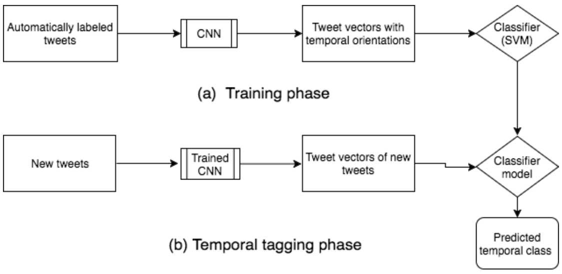
3 Methodology

In this section, we describe our proposed methodology to identify the underlying temporal orientation of tweets and a set of contrastive systems that we used as baselines for comparative study.

3.1 Tweet Temporal Orientation Classifier

The task can be defined as given a tweet t and its posting date d, predict its temporal class $c \in \{$ *past, present*, or *future*$\}$ with reference to its issuing date.

Proposed Architecture: The proposed framework has two main steps as: (i) training the model parameters, and (ii) using the model to tag unseen tweets. During training, we use the weakly labeled tweets to learn the parameters of the CNN and temporal orientation classifier. For classification, a linear Support Vector Machine (lSVM)[1] is used. In particular, we trained three binary classifiers (one per class)[2] using one-vs.-rest, and label a tweet with the class that assigned the highest score. In the second step, we pass tweets through these two optimized components to detect their temporal orientation.


Automatically labeled tweets
CNN
Tweet vectors with temporal orientations
Classifier (SVM)
New tweets
Trained CNN
Tweet vectors of new tweets
Classifier model
Predicted temporal class

(a) Training phase

(b) Temporal tagging phase

Figure 1: Proposed learning architecture.

[1] Trained using the Weka implementation of LIBSVM with linear kernels (polynomial kernels yielded worse performance).

[2] Multi-class classification yielded worse performance.

The choice of CNN for feature extraction is motivated by:

- CNNs have been successfully used as feature extractors in various computer vision tasks and achieved better results compared to handcrafted features. Research has shown that CNN feature maps can be used with SVM to yield classification results that outperform the original CNN (Athiwaratkun et al., 2015)

- Superior accuracies have also been achieved by following a similar line of research in the context of NLP tasks (Kim, 2014; Poria et al., 2015).

Convolutional Neural Networks (CNNs): The task is challenging as tweets are short and noisy. Moreover, English, like many languages, uses a wide variety of ways to refer to the past, present, and future. Unlike previous approaches which mainly rely on hand-crafted rules and feature engineering, we automatically extract features for tweets to build our tweet-level Temporal Orientation Classifier. In particular, we use CNNs to automatically extract tweet vectors as the features for classification. Recently, CNNs have been shown to be useful in many natural language processing and information retrieval tasks by effectively modeling natural language semantics (Collobert et al., 2011). For our experiments, we trained a simple CNN with one convolution layer followed by one max pooling layer (Collobert et al., 2011; Kim, 2014). In the CNN model, we use 3 filters with window sizes of 5, 6 and 7 with 100 feature maps each. These window sizes will capture 5-gram, 6-gram, and 7-gram information in the tweets. We employ dropout for regularization with a dropout rate of 0.5 which is a reasonable default. We also use rectified linear units and mini-batches with size of 50. The parameters of the CNN were fixed based on the performance of 3-fold cross-validation. The tweet representations are trained on top of pre-trained word vectors which are updated during CNN training. We use the publicly available word2vec[3] vectors that are trained on Google News corpus as well our own Word2vec vectors[4] trained during the labeled-data creation phase. During the training phase, the parameters of the CNN model are learned by passing multiple filters over word vectors and then applying the max-over-time pooling operation to generate features which are used in a fully connected softmax layer. Finally, we use the cross-entropy loss function for learning the parameters of the model. Similar to Kim (2014), we use dropout (Hinton et al., 2012) to regularize the change of parameters by randomly setting some weights to zero that prevents overfitting.

3.2 Income Predictor Model

Similar to Preoţiuc-Pietro et al. (2015), we formulate the income prediction task as regression using user-level temporal orientation as features. First, the tweet temporal orientation classifier is used to label whether a tweet focuses on past, present, or future. Afterwards, at user-level, we produce three categories of temporal orientation (three separate variables summing to one), defined simply as the proportion of a user's total tweets ($tweets(user)_{all}$) classified in the given temporal category ($c \in \{$ *past, present*, or *future*$\}$), as in (1):

$$orientation_c(user) = \frac{|tweets_c(user)|}{|tweets_{all}(user)|} \quad (1)$$

We use linear and non-linear methods. The linear method is logistic regression (LR) (Freedman, 2009) with Elastic Net regularisation. In order to capture the non-linear relationship between a user's temporal orientation and their income, we use Gaussian Processes (GP) (Rasmussen and Nickisch, 2010) for regression. Given that our dataset is very large and the number of features is high, for GP inference we use the fully independent training conditional approximation (Snelson and Ghahramani, 2005) with 500 random inducing points.

4 Data Sets

4.1 Training Data

Tweets are collected using the Twitter streaming API.[5] We downloaded English tweets during the period 01.01.2015–31.01.2015, which generated about 40 million tweets. After collecting the tweets, we filter past-, present-, and future-oriented tweets using a manually selected high precision list of 50 seed terms. These are terms

[3] https://code.google.com/p/word2vec/

[4] trained using gensim library available at https://radimrehurek.com/gensim/intro.html

[5] https://dev.twitter.com/streaming/overview.

that capture temporal dimensions of tweets with very few false positives, though the recall of these terms is low. In order to increase the recall, and to capture new terms that are good paradigms of past, present, and future, we expand our initial seed terms using a query expansion technique. We employ a continuous distributed vector representation of words using the continuous Skip-gram model (also known as Word2Vec) proposed by Mikolov et al. (2013). The model is trained on the whole collection of 40 million tweets with dimension and window size set to 300 and 7, respectively.

Given the vector representations for the terms, we calculate the similarity scores between pairs of terms in our vocabulary using cosine similarity. The top 10 similar terms for each seed term are selected for the expansion of the initial seed list. We again filter the whole collection of tweets using the newly added seed terms. We finally select 120,000 tweets equally distributed in past (=40,000 tweets), present (=40,000), and future (=40,000) temporal categories.[6] Examples of filtered tweets are as follows:

- *Thank you so much for coming in for our show yesterday.* (**seed=yesterday**)
- *@**** is currently out of the office working his other job.* (**seed=currently**)
- *I promise you don't have to be afraid.* (**seed=promise**)

Table 1 shows some examples of expanded terms for some of the initial seed terms. There are some unrelated keywords in the expanded seed list due to the automatic process of keyword selection.

4.2 Test Set

In order to evaluate the tweet temporal orientation classification model, 2035 tweets were manually annotated by three human annotators in four different categories: past, present, future and doubtful. Majority voting is applied to assign the final output class to a given tweet. Tweets whose temporal orientation was not resolved by majority voting were deleted from the test set.[7] The final distribution of annotated tweets was: past=423, present=1252, future=325, doubtful=35.

4.3 Income data of Users

We used a dataset developed by Preoţiuc-Pietro et al. (2015), which contains 5,191 Twitter users along with their platform statistics and ≈10 million historical tweets. The dataset is based on mapping a Twitter user to a job title and using this as a proxy for the mean income for that specific occupation.

5 Experimental Results

Temporal Orientation Classification Results: The performance of our tweet temporal orientation classifier is evaluated using the manually annotated test set. We compare our approach with two baselines that are the most relevant for our research: (i) Baseline1: a rule-based method (Nie et al., 2015) and (ii) Baseline2: a supervised learning strategy with bag-of-words, time expressions, part-of-speech tags, and temporal class-specific lexicon features (Schwartz et al., 2015). Comparative evaluation results are presented in Table 2. The results show that our weakly supervised framework outperforms rule-based and supervised learning technique in terms of accuracy.

We examine the impact of the size of labeled training data on each method's performance. Baseline1 (rule-based approach) is not involved since this does not depend on labeled training data. We randomly select $d\%$ of the training data to train the classifiers and test them on the test set, with d ranging from 10 to 90. For each d, we generate the training set 20 times and the averaged performance is recorded. Accuracies of both approaches over the test data are presented in Table 3. Results show that our proposed framework performs consistently better than its counterpart. In particular, results show that with 30K training examples, better results can be obtained by our approach than relying on 120K training items for the state-of-the-art supervised machine learning approach (Baseline2).

Income Prediction Results: Similar to Preoţiuc-Pietro et al. (2015), we measure the predictive power of temporal orientation by performing regression on the user income. Performance is measured using 10-fold cross-validation: in each round, 80% of the data is used to train the model, 10% is used to tune model parameters using grid search and a different 10% is held out for testing. The final results are computed over the aggregate set of results of all 10 folds. Results us-

[6] Similar to Schwartz et al. (2015), we only considered past, present and future categories.

[7] Note that we approached the authors of Schwartz et al. (2015) to obtain their dataset but they did not share the data because of copyright issues. This is the reason for generating our own gold-standard test set.

Initial Seed Terms (Temporal Orientation)	**Extended Seed Terms**
Yesterday (Past)	yesterday!, started, yday, finished, already, yest, earlier, held, arrived
Currently (Present)	now, still, presently, available, whilst, actively, contemplating, considering
Promise (Future)	guarantee, expect, doubt, commitment, think, hope, opportunity, tomorrow

Table 1: Examples of initial seed terms and expanded seed terms.

Method	**Baseline1**	**Baseline2**	**Proposed Method**[1]	**Proposed Method**[2]
Accuracy	48.8	67.4	**74.4**	72.7
Past (p, r, f1)	(52.0, 56.3, 54.0)	(67.4, 81.9, 73.9)	(84.5, 79.8, 82.0)	(71.1, 79.5, 75.0)
Present (p, r, f1)	(58.2, 54.2, 56.1)	(69.3, 82.6, 75.3)	(81.3, 86.6, 83.8)	(73.0, 71.5, 72.2)
Future (p, r, f1)	(51.0, 53.3, 52.1)	(64.4, 77.9, 70.5)	(78.5, 79.8, 79.1)	(79.4, 69.5, 74.0)

Table 2: Accuracy for *past, present, future* classifications using different methods measured over test data. Results are broken down by precision (p), recall (r), and f1-measure (f1) scores. Proposed Method[1] and Proposed Method[2] represent our classification framework with Word2vec vectors derived from our collected tweet and pre-trained Google News corpus, respectively.

Training data size	**Baseline2**	**Proposed Method**[1]
10k	57.5	61.3
20k	60.2	66.4
30k	63.5	71.7
50k	65.4	73.6
70k	66.1	74.2
90k	67.4	74.1
120K (all)	67.4	**74.4**

Table 3: Tweets temporal orientation classification accuracies with different sizes of training data.

ing linear and non-linear regression methods and past, present, future temporal orientation features are presented in Table4. Performance is measured using two standard metrics: Pearson's correlation coefficient r and Mean Absolute Error (MAE) between inferred and target values. Results show that correlation between a user's future temporal orientation and their income is the highest, i.e. people with higher future temporal orientation tend to have higher income levels. Results also demonstrate that predictive models with future temporal orientation as a feature can predict income with high accuracy compared to past and present temporal orientation. Our findings are consistent with previous research that suggests that future-oriented thinking is linked to academic achievement, increased social involvement, lower distress, extroversion, and conscientiousness. These factors are also positively correlated with income (Kahana et al., 2005; Roberts et al., 2007). Note also that, the non-linear methods outperform the linear methods by a wide margin, showing the importance of modeling non-linear relationships in our data.

Method	Temporal Orientation	Correlation coefficient	MAE
LR	Past Present Future	0.1449 0.0998 **0.4505**	£12365 £14365 £ 10850
GP	Past Present Future	0.1849 0.1099 **0.5104**	£11200 £12125 £ 10235

Table 4: Prediction of income using temporal orientation features

6 Conclusions

We presented the first large-scale study aiming to predict the income of Twitter users from their temporal orientation. Temporal orientation of users is assessed from their tweets. Our weakly supervised learning framework automatically time-tags tweets according to its underlying temporal orientation: past, present, or future. The associations we found between user-level future temporal orientation and income are novel in the context of well-established temporal orientation correlates. As future work, we are in the process of improving the temporal orientation classification accuracy by incorporating linguistic and sentiment related features into the deep learning phase.

Acknowledgments

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Jean Adams and Daniel Nettle. 2009. Time perspective, personality and smoking, body mass, and physical activity: An empirical study. *British journal of health psychology* 14(1):83–105.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Paris, France, April 27 - May 2, 2013*. pages 3267–3276.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one* 6(12):e26752.

David A Freedman. 2009. *Statistical models: theory and practice*. Cambridge University Press, Cambridge, UK.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Massachusetts, US. http://www.deeplearningbook.org.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* .

Eva Kahana, Boaz Kahana, and Jianping Zhang. 2005. Motivational antecedents of preventive proactivity in late life: Linking future orientation and exercise. *Motivation and emotion* 29(4):438–459.

Daniel Kahneman and Angus Deaton. 2010. High income improves evaluation of life but not emotional well-being. *Proceedings of the national academy of sciences* 107(38):16489–16493.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110(15):5802–5805.

James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. 2014. Age and gender identification in social media. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*. pages 1129–1136.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems, Lake Tahoe, Nevada, United States, December 5-8, 2013*. pages 3111–3119.

Aiming Nie, Jason Shepard, Jinho Choi, Bridget Copley, and Phillip Wolff. 2015. Computational exploration of the linguistic structures of future-oriented expression: Classification and categorization. In *NAACL-HLT 2015 Student Research Workshop (SRW), Denver, Colorado, USA*. volume 867, page 168.

Gregory Park, H. Andrew Schwartz, Maarten Sap, Margaret L. Kern, Evan Weingarten, Johannes C. Eichstaedt, Jonah Berger, David J. Stillwell, Michal Kosinski, Lyle H. Ungar, and Martin E. P. Seligman. 2017. Living in the past, present, and future: Measuring temporal orientation with language. *Journal of Personality* 85(2):270–280.

Soujanya Poria, Erik Cambria, and Alexander F. Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. pages 2539–2544.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PloS one* 10(9):e0138717.

Carl Edward Rasmussen and Hannes Nickisch. 2010. Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research* 11:3011–3015.

Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. 2007. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science* 2(4):313–345.

Maarten Sap, Gregory J. Park, Johannes C. Eichstaedt, Margaret L. Kern, David Stillwell, Michal Kosinski, Lyle H. 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, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 1146–1151.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al.

2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791.

H. Andrew Schwartz, Greg Park, Maarten Sap, Evan Weingarten, Johannes Eichstaedt, Margaret Kern, Jonah Berger, Martin Seligman, and Lyle Ungar. 2015. Extracting human temporal orientation in facebook language. In *Proceedings of The 2015 Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies (NAACL), Denver, Colorado, USA*. pages 409–419.

Edward Snelson and Zoubin Ghahramani. 2005. Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*. pages 1257–1264.

Paul Webley and Ellen K Nyhus. 2006. Parents influence on childrens future orientation and saving. *Journal of Economic Psychology* 27(1):140–164.

Philip G Zimbardo and John N Boyd. 2015. Putting time in perspective: A valid, reliable individual-differences metric. In *Time Perspective Theory; Review, Research and Application*, Springer, Berlin, Germany, pages 17–55.