

# Predicting the Industry of Users on Social Media

Konstantinos Pappas  
University of Michigan  
Computer Science and Engineering  
Ann Arbor, MI 48109, USA  
pappus@umich.edu

Rada Mihalcea  
University of Michigan  
Computer Science and Engineering  
Ann Arbor, MI 48109, USA  
mihalcea@umich.edu

## ABSTRACT

Automatic profiling of social media users is an important task for supporting a multitude of downstream applications. While a number of studies have used social media content to extract and study collective social attributes, there is a lack of substantial research that addresses the detection of a user’s industry. We frame this task as classification using both feature engineering and ensemble learning. Our industry-detection system uses both posted content and profile information to detect a user’s industry with 64.3% accuracy, significantly outperforming the majority baseline in a taxonomy of fourteen industry classes. Our qualitative analysis suggests that a person’s industry not only affects the words used and their perceived meanings, but also the number and type of emotions being expressed.

## CCS Concepts

•**Social and professional topics** → **User characteristics**; •**Human-centered computing** → *Social media*; •**Applied computing** → *Document analysis*; *Sociology*;

## Keywords

User Profiling; Social Media; Sociolinguistics

## 1. INTRODUCTION

Over the past two decades, the emergence of social media has enabled the proliferation of traceable human behavior. The content posted by users can reflect who their friends are, what topics they are interested in, or which company they are working for. At the same time, users are listing a number of profile fields to define themselves to others. The utilization of such metadata has proven important in facilitating further developments of applications in advertising [6], personalization [9], and recommender systems [1]. However, profile information can be limited, depending on the platform, or it is often deliberately omitted [18]. To uncloak this information, a number of studies have utilized social media users’ footprints to approximate their profiles.

This paper explores the potential of predicting a user’s industry –the aggregate of enterprises in a particular field– by identifying industry indicative text in social media. The accurate prediction of users’ industry can have a big impact on targeted advertising by minimizing wasted advertising [20] and improved personalized user experience. A number of studies in the social sciences have associated

language use with social factors such as occupation, social class, education, and income [4, 23, 5, 24]. An additional goal of this paper is to examine such findings, and in particular the link between language and occupational class, through a data-driven approach.

In addition, we explore how meaning changes depending on the occupational context. By leveraging word embeddings, we seek to quantify how, for example, *cloud* might mean a separate concept (e.g., condensed water vapor) in the text written by users that work in environmental jobs while it might be used differently by users in technology occupations (e.g., Internet-based computing).

Specifically, this paper makes four main contributions. First, we build a large, industry-annotated dataset that contains over 20,000 blog users. In addition to their posted text, we also link a number of user metadata including their gender, location, occupation, introduction and interests.

Second, we build content-based classifiers for the industry prediction task and study the effect of incorporating textual features from the users’ profile metadata using various meta-classification techniques, significantly improving both the overall accuracy and the average per industry accuracy.

Next, after examining which words are indicative for each industry, we build vector-space representations of word meanings and calculate one deviation for each industry, illustrating how meaning is differentiated based on the users’ industries. We qualitatively examine the resulting industry-informed semantic representations of words by listing the words per industry that are most similar to job related and general interest terms.

Finally, we rank the different industries based on the normalized relative frequencies of emotionally charged words (positive and negative) and, in addition, discover that, for both genders, these frequencies do not statistically significantly correlate with an industry’s gender dominance ratio.

After discussing related work in Section 2, we present the dataset used in this study in Section 3. In Section 4 we evaluate two feature selection methods and examine the industry inference problem using the text of the users’ postings. We then augment our content-based classifier by building an ensemble that incorporates several metadata classifiers. We list the most industry indicative words and expose how each industrial semantic field varies with respect to a variety of terms in Section 5. We explore how the frequencies of emotionally charged words in each gender correlate with the industries and their respective gender dominance ratio and, finally, conclude in Section 6.

Technology	4,175	Law	1,520
Religion	3,165	Security/Military	933
Fashion	2,119	Tourism	840
Publishing	2,102	Construction	837
Sports or Recreation	1,779	Museums or Libraries	823
Real Estate	1,726	Banking/Investment Banking	735
Agriculture/Environment	1,620	Automotive	506

**Table 1: Industry categories and number of users per category.**

Data per User	max	mean	$\sigma$	median
Blogs	97	1.8	2.9	1
Blog Posts	1356	24.5	30.4	21
Characters	4,939,258	56,948	112,048.1	33,404

**Table 2: Statistics on the Blogger dataset.**

## 2. RELATED WORK

Alongside the wide adoption of social media by the public, researchers have been leveraging the newly available data to create and refine models of users’ behavior and profiling. There exists a myriad research that analyzes language in order to profile social media users. Some studies sought to characterize users’ personality [30, 7], while others sequenced the expressed emotions [13], studied mental disorders [8], and the progression of health conditions [21]. At the same time, a number of researchers sought to predict the social media users’ age and/or gender [34, 33, 22], while others targeted and analyzed the ethnicity, nationality, and race of the users [14, 32, 29]. One of the profile fields that has drawn a great deal of attention is the location of a user. Among others, Hecht et al. [17] predicted Twitter users’ locations using machine learning on nationwide and state levels. Later, Han et al. [16] identified location indicative words to predict the location of Twitter users down to the city level.

As a separate line of research, a number of studies have focused on discovering the political orientation of users [33, 25, 37]. Finally, Li et al. [26] proposed a way to model major life events such as getting married, moving to a new place, or graduating. In a subsequent study, [27] described a weakly supervised information extraction method that was used in conjunction with social network information to identify the name of a user’s spouse, the college they attended, and the company where they are employed.

The line of work that is most closely related to our research is the one concerned with understanding the relation between people’s language and their industry. Previous research from the fields of psychology and economics have explored the potential for predicting one’s occupation from their ability to use math and verbal symbols [11] and the relationship between job-types and demographics [35]. More recently, Huang et al. [19] used machine learning to classify Sina Weibo users to twelve different platform-defined occupational classes highlighting the effect of homophily in user interactions. This work examined only users that have been verified by the Sina Weibo platform, introducing a potential bias in the resulting dataset. Finally, Preotiuc-Pietro et al. [31] predicted the occupational class of Twitter users using the Standard Occupational Classification (SOC) system, which groups the different jobs based on skill requirements. In that work, the data collection process was limited to only users that specifically mentioned their occupation in their self-description in a way that could be directly mapped to a

SOC occupational class. The mapping between a substring of their self-description and a SOC occupational class was done manually. Because of the manual annotation step, their method was not scalable; moreover, because they identified the occupation class inside a user self-description, only a very small fraction of the Twitter users could be included (in their case, 5,191 users).

Both of these recent studies are based on micro-blogging platforms, which inherently restrict the number of characters that a post can have, and consequently the way that users can express themselves.

Moreover, both studies used off-the-shelf occupational taxonomies (rather than self-declared occupation categories), resulting in classes that are either too generic (e.g., media, welfare and electronic are three of the twelve Sina Weibo categories), or too intermixed (e.g., an assistant accountant is in a different class from an accountant in SOC). To address these limitations, we investigate the industry prediction task in a large blog corpus consisting of over 20K American users, 40K web-blogs, and 560K blog posts.

## 3. DATASET

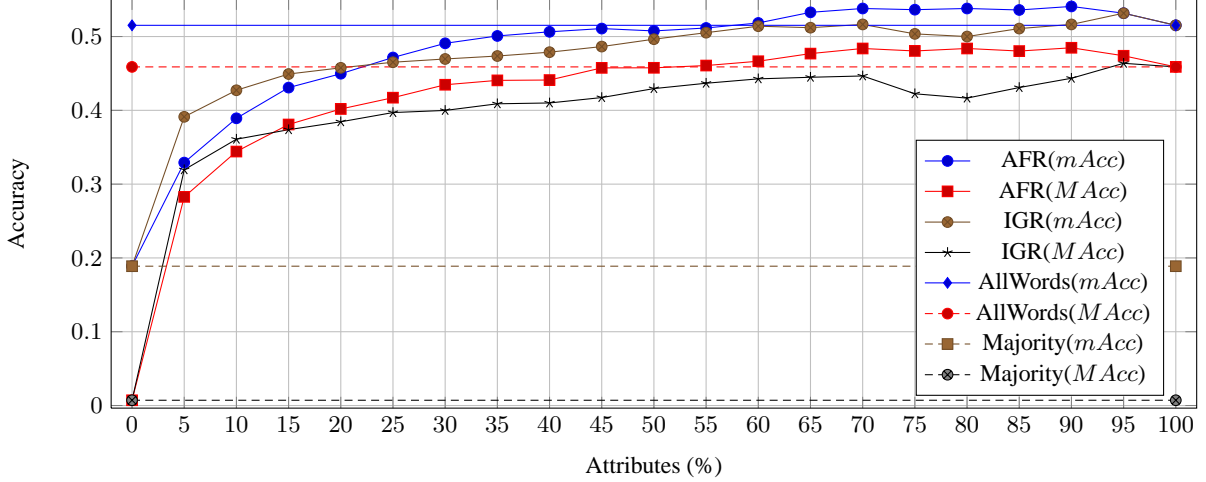
We compile our industry-annotated dataset by identifying blogger profiles located in the U.S. on the profile finder on <http://www.blogger.com>, and scraping only those users that had the industry profile element completed.<sup>1</sup>

For each of these bloggers, we retrieve all their blogs, and for each of these blogs we download the 21 most recent blog postings. We then clean these blog posts of HTML tags and tokenize them, and drop those bloggers whose cumulative textual content in their posts is less than 600 characters. Following these guidelines, we identified all the U.S. bloggers with completed industry information.

Traditionally, standardized industry taxonomies organize economic activities into groups based on similar production processes, products or services, delivery systems or behavior in financial markets. Following such assumptions and regardless of their many similarities, a tomato farmer would be categorized into a distinct industry from a tobacco farmer. As demonstrated in Preotiuc-Pietro et al. [31] such groupings can cause unwarranted misclassifications.

The Blogger platform provides a total of 39 different industry options. Even though a completed industry value is an implicit text

<sup>1</sup>This data collection was performed between May and July 2015.



**Figure 1: Feature evaluation on the industry prediction task using Information Gain Ratio (IGR) and our Aggressive Feature Ranking (AFR). The performance is measured using both accuracy ( $mAcc$ ) and average per-class accuracy ( $MAcc$ ).**

annotation, we acknowledge the same problem noted in previous studies: some categories are too broad, while others are very similar. To remedy this and following Guibert et al. [15], who argued that the denominations used in a classification must reflect the purpose of the study, we group the different Blogger industries based on similar educational background and similar technical terminology. To do that, we exclude very general categories and merge conceptually similar ones<sup>2</sup>. Examples of broad categories are the *Education* and the *Student* options: a teacher could be teaching in any concentration, while a student could be enrolled in any discipline. Examples of conceptually similar categories are the *Investment Banking* and the *Banking* options.

The final set of categories is shown in Table 1, along with the number of users in each category. The resulting dataset consists of 22,880 users, 41,094 blogs, and 561,003 posts. Table 2 presents additional statistics of our dataset.

## 4. TEXT-BASED INDUSTRY MODELING

After collecting our dataset, we split it into three sets: a train set, a development set, and a test set. The sizes of these sets are 17,880, 2,500, and 2,500 users, respectively, with users randomly assigned to these sets. In all the experiments that follow, we evaluate our classifiers by training them on the train set, configure the parameters and measure performance on the development set, and finally report the prediction accuracy and results on the test set. Note that all the experiments are performed at user level, i.e., all the data for one user is compiled into one instance in our data sets.

To measure the performance of our classifiers, we use the prediction accuracy. However, as shown in Table 1, the available data is skewed across categories, which could lead to somewhat distorted accuracy numbers depending on how well a model learns to predict the most populous classes. Moreover, accuracy alone does not provide a great deal of insight into the individual performance per industry, which is one of the main objectives in this study. Therefore, in our results below, we report: (1) micro-accuracy ( $mAcc$ ),

<sup>2</sup>Merged categories are denoted with the ‘/’ character in Table 1.

calculated as the percentage of correctly classified instances out of all the instances in the development (test) data; and (2) macro-accuracy ( $MAcc$ ), calculated as the average of the per-category accuracies, where the per-category accuracy is the percentage of correctly classified instances out of the instances belonging to one category in the development (test) data.

### 4.1 Leveraging Blog Content

In this section, we seek the effectiveness of using solely textual features obtained from the users’ postings to predict their industry.

The industry prediction baseline *Majority* is set by discovering the most frequently featured class in our training set and picking that class in all predictions in the respective development or testing set.

After excluding all the words that are not used by at least three separate users in our training set, we build our *AllWords* model by counting the frequencies of all the remaining words and training a multinomial Naive Bayes classifier. As seen in Figure 1, we can far exceed the *Majority* baseline performance by incorporating basic language signals into machine learning algorithms (173%  $mAcc$  improvement).

We additionally explore the potential of improving our text classification task by applying a number of feature ranking methods and selecting varying proportions of top ranked features in an attempt to exclude noisy features. We start by ranking the different features,  $w$ , according to their Information Gain Ratio score (IGR) with respect to every industry,  $i$ , and training our classifier using different proportions of the top features.

$$IGR(w) = \frac{IG(w)}{IV(w)} \propto \frac{-H(i|w)}{-P(w)\log P(w) - P(\bar{w})\log P(\bar{w})} \propto$$

$$\frac{P(w) \sum_{i \in I} P(i|w) \log P(i|w) + P(\bar{w}) \sum_{i \in I} P(i|\bar{w}) \log P(i|\bar{w})}{-P(w)\log P(w) - P(\bar{w})\log P(\bar{w})}$$

Even though we find that using the top 95% of all the features al-

Data	Gender	Occupation	City	State	Introduction	Interests
Train	0.806	0.753	0.862	1.00	0.692	0.535
Dev	0.814	0.712	0.788	1.00	0.671	0.549
Test	0.812	0.709	0.768	1.00	0.686	0.533

**Table 3: Proportion of users with non-empty metadata fields.**

ready exceeds the performance of the *All Words* model on the development data, we further experiment with ranking our features with a more aggressive formula that heavily promotes the features that are tightly associated with any industry category. Therefore, for every word in our training set, we define our newly introduced ranking method, the Aggressive Feature Ranking (AFR), as:

$$AFR(w) = \max_{i \in I} \frac{P(w|i)}{P(w)}$$

In Figure 1 we illustrate the performance of all four methods in our industry prediction task on the development data. Note that for each method, we provide both the accuracy (*mAcc*) and the average per-class accuracy (*MAcc*). The *Majority* and *All Words* methods apply to all the features; therefore, they are represented as a straight line in the figure. The *IGR* and *AFR* methods are applied to varying subsets of the features using a 5% step.

Our experiments demonstrate that the word choice that the users make in their posts correlates with their industry. The first observation in Figure 1 is that the *mAcc* is proportional to *MAcc*; as *mAcc* increases, so does *MAcc*. Secondly, the best result on the development set is achieved by using the top 90% of the features using the *AFR* method. Lastly, the improvements of the *IGR* and *AFR* feature selections are not substantially better in comparison to *All Words* (at most 5% improvement between *All Words* and *AFR*), which suggest that only a few noisy features exist and most of the words play some role in shaping the “language” of an industry.

As a final evaluation, we apply on the test data the classifier found to work best on the development data (*AFR* feature selection, top 90% features), for an *mAcc* of 0.534 and *MAcc* of 0.477.

## 4.2 Leveraging User Metadata

Together with the industry information and the most recent postings of each blogger, we also download a number of accompanying profile elements. Using these additional elements, we explore the potential of incorporating users’ metadata in our classifiers.

Table 3 shows the different user metadata we consider together with their coverage percentage (not all users provide a value for all of the profile elements). With the exception of the gender field, the remaining metadata elements shown in Table 3 are completed by the users as a freely editable text field. This introduces a considerable amount of noise in the set of possible metadata values. Examples of noise in the occupation field include values such as “Retired”, “I work.”, or “momma” which are not necessarily informative for our industry prediction task.

To examine whether the metadata fields can help in the prediction of a user’s industry, we build classifiers using the different metadata elements. For each metadata element that has a textual value, we use all the words in the training set for that field as features. The only two exceptions are the *state* field, which is encoded as one

Classifier	<i>mAcc</i>	<i>MAcc</i>
OCCU	<b>0.566</b>	<b>0.431</b>
INTRO	0.406	0.247
INTER	0.287	0.157
GLOC	0.199	0.090

**Table 4: Accuracy (*mAcc*) and average per-class accuracy (*MAcc*) of the base metadata classifiers on the development set.**

TEXT	0.245	0.338	0.357	0.366
0.270	OCCU	0.348	0.386	0.409
0.186	0.303	TEXT	0.535	0.554
0.019	0.153	0.216	INTER	0.668
-0.129	-0.005	0.005	0.020	GLOC

**Table 5: Kappa scores and double fault results of the base classifiers on development data.**

feature that can take one out of 50 different values representing the 50 U.S. states; and the *gender* field, which is encoded as a feature with a distinct value for each user gender option: undefined, male, or female.

As shown in Table 4, we build four different classifiers using the multinomial NB algorithm: OCCU (which uses the words found in the *occupation* profile element), INTRO (*introduction*), INTER (*interests*), and GLOC (combined *gender*, *city*, *state*).

In general, all the metadata classifiers perform better than our majority baseline (*mAcc* of 18.88%). For the GLOC classifier, this result is in alignment with previous studies [35]. However, the only metadata classifier that outperforms the content classifier is the OCCU classifier, which despite missing and noisy *occupation* values exceeds the content classifier’s performance by an absolute 3.2%.

To investigate the promise of combining the five different classifiers we have built so far, we calculate their inter-prediction agreement using Fleiss’s Kappa [10], as well as the lower prediction bounds using the double fault measure [12]. The Kappa values, presented in the lower left side of Table 5, express the classification agreement for categorical items, in this case the users’ industry. Lower values, especially values below 30%, mean smaller agreement. Since all five classifiers have better-than-baseline accuracy, this low agreement suggests that their predictions could potentially be combined to achieve a better accumulated result.

Moreover, the double fault measure values, which are presented in the top-right hand side of Table 5, express the proportion of test cases for which both of the two respective classifiers make false predictions, essentially providing the lowest error bound for the pairwise ensemble classifier performance. The lower those num-

Meta-classifiers	Feature Concatenation		Stacking	
	$mAcc$	$MAcc$	$mAcc$	$MAcc$
1. TEXT + OCCU	0.545	0.489	0.640	0.557
2. {1} + INTRO	0.546	0.487	0.648	0.560
3. {2} + INTER	0.546	0.482	<b>0.653</b>	<b>0.569</b>
4. {3} + GLOC	0.545	0.478	0.650	0.566

**Table 6: Performance of feature concatenation (early fusion) and stacking (late fusion) on the development set.**

bers are, the greater the accuracy potential of any meta-classification scheme that combines those classifiers. Once again, the low double fault measure values suggest potential gain from a combination of the base classifiers into an ensemble of models.

After establishing the promise of creating an ensemble of classifiers, we implement two meta-classification approaches. First, we combine our classifiers using features concatenation (or early fusion). Starting with our content-based classifier (TEXT), we successively add the features derived from each metadata element. The results, both micro- and macro-accuracy, are presented in Table 6. Even though all these four feature concatenation ensembles outperform the content-based classifier in the development set, they fail to outperform the OCCU classifier.

Second, we explore the potential of using stacked generalization (or late fusion) [38]. The base classifiers, referred to as L0 classifiers, are trained on different folds of the training set and used to predict the class of the remaining instances. Those predictions are then used together with the true label of the training instances to train a second classifier, referred to as the L1 classifier: this L1 is used to produce the final prediction on both the development data and the test data. Traditionally, stacking uses different machine learning algorithms on the same training data. However in our case, we use the same algorithm (multinomial NB) on heterogeneous data (i.e., different types of data such as content, occupation, introduction, interests, gender, city and state) in order to exploit all available sources of information.

The ensemble learning results on the development set are shown in Table 6. We notice a constant improvement for both metrics when adding more classifiers to our ensemble except for the GLOC classifier, which slightly reduces the performance. The best result is achieved using an ensemble of the TEXT, OCCU, INTRO, and INTER L0 classifiers; the respective performance on the test set is an  $mAcc$  of 0.643 and an  $MAcc$  of 0.564.

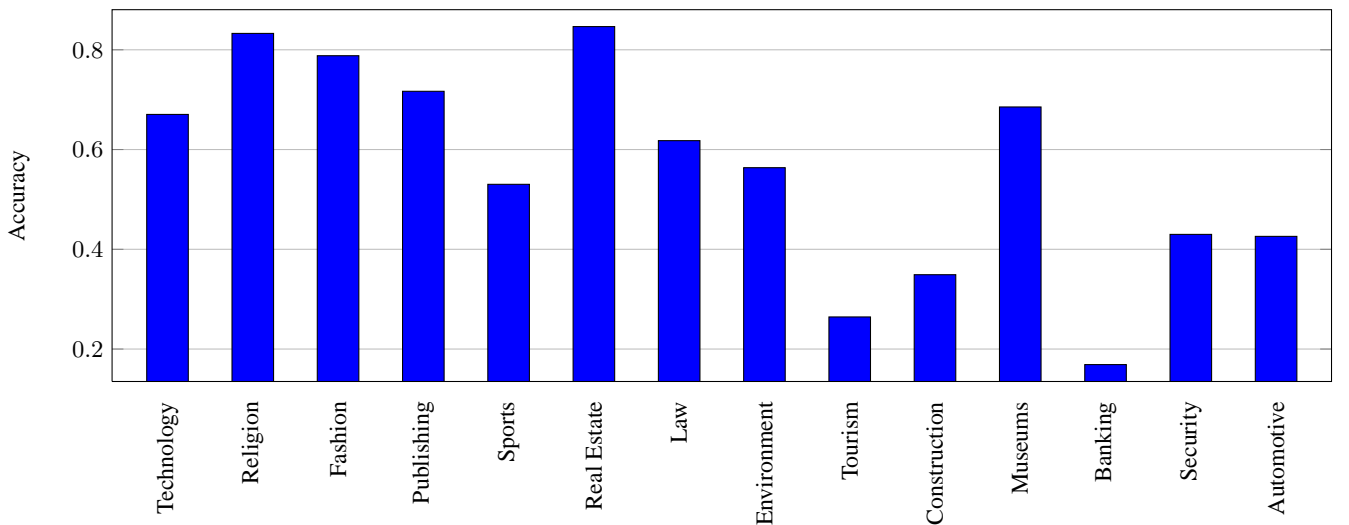
Finally, we present in Figure 2 the prediction accuracy for the final classifier for each of the different industries in our test dataset. Evidently, some industries are easier to predict than others. For example, while the *Real Estate* and *Religion* industries achieve accuracy figures above 80%, other industries, such as the *Banking* industry, are predicted correctly in less than 17% of the time. Anecdotal evidence drawn from the examination of the confusion matrix does not encourage any strong association of the *Banking* class with any other. The misclassifications are roughly uniform across all other classes, suggesting that the users in the *Banking* industry use language in a non-distinguishing way.

## 5. QUALITATIVE ANALYSIS

In this section, we provide a qualitative analysis of the language of the different industries.

### 5.1 Top-Ranked Words

To conduct a qualitative exploration of which words indicate the in-



**Figure 2: Accuracy per-class using stacking meta-classification.**

dustry of a user, Table 7 shows the three top-ranking content words for the different industries using the *AFR* method.

Industry	Top-Ranked Words
Technology	software, file, data
Religion	ministry, jesus, pastor
Fashion	fashion, dress, hair
Publishing	writers, novel, writer
Sports	coach, weight, exercise
Real Estate	estate, details, homes
Law	court, trial, agreement
Environment	farm, plants, plant
Tourism	guests, travel, hotel
Construction	roof, construction, union
Museums	library, museum, novel
Banking	secret, agent, bank
Security	officer, army, military
Automotive	vehicle, cars, insurance

**Table 7: Three top-ranked words for each industry.**

Not surprisingly, the top ranked words align well with what we would intuitively expect for each industry. Even though most of these words are potentially used by many users regardless of their industry in our dataset, they are still distinguished by the *AFR* method because of the different frequencies of these words in the text of each industry.

## 5.2 Industry-specific Word Similarities

Next, we examine how the meaning of a word is shaped by the context in which it is uttered. In particular, we qualitatively investigate how the speakers’ industry affects meaning by learning vector-space representations of words that take into account such contextual information. To achieve this, we apply the contextualized word embeddings proposed by Bamman et al. [2], which are based on an extension of the “skip-gram” language model [28].

Technology		Tourism	
term	cosine	term	cosine
customers	1.000	customers	1.000
clients	0.870	guests	0.816
consumers	0.858	opportunities	0.789
companies	0.832	clients	0.783
employees	0.822	itineraries	0.778
users	0.820	choices	0.769
developers	0.818	patrons	0.767
providers	0.817	employees	0.760
businesses	0.813	projects	0.757
customer	0.811	provide	0.753

**Table 8: Terms with the highest cosine similarity to the term *customers*.**

In addition to learning a global representation for each word, these contextualized embeddings compute one deviation from the common word embedding representation for each contextual variable, in this case, an industry option. These deviations capture the terms’ meaning variations (shifts in the  $k$ -dimensional space of the representations, where  $k = 100$  in our experiments) in the text of the different industries, however all the embeddings are in the same vector space to allow for comparisons to one another.

Environment		Tourism	
term	cosine	term	cosine
food	1.000	food	1.000
local	0.824	delicious	0.843
produce	0.812	treats	0.822
meat	0.807	pastries	0.814
wholesome	0.805	sandwiches	0.808
processed	0.785	burgers	0.806
consumers	0.777	dishes	0.801
meals	0.774	selections	0.796
nutritionally	0.774	eating	0.792
locally	0.765	hamburgers	0.791

**Table 9: Terms with the highest cosine similarity to the term *food*.**

Using the word representations learned for each industry, we present in Table 8 the terms in the *Technology* and the *Tourism* industries that have the highest cosine similarity with a job-related word, *customers*. Similarly, Table 9 shows the words in the *Environment* and the *Tourism* industries that are closest in meaning to a general interest word, *food*. More examples are given in the Appendix A.

The terms that rank highest in each industry are noticeably different. For example, as seen in Table 9, while *food* in the *Environment* industry is similar to *nutritionally* and *locally*, in the *Tourism* industry the same word relates more to terms such as *delicious* and *pastries*. These results not only emphasize the existing differences in how people in different industries perceive certain terms, but they also demonstrate that those differences can effectively be captured in the resulting word embeddings.

## 5.3 Emotional Orientation per Industry and Gender

As a final analysis, we explore how words that are emotionally charged relate to different industries. To quantify the emotional orientation of a text, we use the *Positive Emotion* and *Negative Emotion* categories in the Linguistic Inquiry and Word Count (LIWC) dictionary [36]. The LIWC dictionary contains lists of words that have been shown to correlate with the psychological states of people that use them; for example, the *Positive Emotion* category contains words such as “happy,” “pretty,” and “good.”

For the text of all the users in each industry we measure the frequencies of *Positive Emotion* and *Negative Emotion* words normalized by the text’s length. Table 10 presents the industries’ ranking for both categories of words based on their relative frequencies in the text of each industry.

We further perform a breakdown per-gender, where we once again calculate the proportion of emotionally charged words in each industry, but separately for each gender. We find that the industry rankings of the relative frequencies  $f_i$  of emotionally charged words for the two genders are statistically significantly correlated,<sup>3</sup> which suggests that regardless of their gender, users use positive (or negative) words with a relative frequency that correlates with their industry. (In other words, even if e.g., *Fashion* has a larger number of women users, both men and women working in *Fashion* will tend to use more positive words than the corresponding gender in another industry with a larger number of men users such as *Automotive*.)

<sup>3</sup> $\rho > 0.81$  and  $p < 0.001$  for both categories of words.

Positive	$f_i \times 10^3$	Negative	$f_i \times 10^3$
Fashion	35.93	Security	13.80
Religion	32.10	Religion	13.68
Tourism	30.61	Law	12.97
Banking	30.44	Publishing	12.66
Sports	30.05	Construction	11.77
Real Estate	29.25	Banking	11.74
Publishing	29.12	Sports	10.68
Security	28.92	Technology	10.65
Construction	28.84	Museums	10.55
Museums	28.82	Automotive	10.53
Environment	28.31	Environment	10.17
Law	27.63	Tourism	9.53
Automotive	27.17	Fashion	8.50
Technology	26.42	Real Estate	8.25

**Table 10: Ranking of industries based on the relative frequencies  $f_i$  of Positive Emotion and Negative Emotion words.**

Finally, motivated by previous findings of correlations between job satisfaction and gender dominance in the workplace [3], we explore the relationship between the usage of *Positive Emotion* and *Negative Emotion* words and the gender dominance in an industry. Although we find that there are substantial gender imbalances in each industry (Appendix B), we did not find any statistically significant correlation between the gender dominance ratio in the different industries and the usage of positive (or negative) emotional words in either gender in our dataset.

## 6. CONCLUSION

In this paper, we examined the task of predicting a social media user’s industry. We introduced an annotated dataset of over 20,000 blog users and applied a content-based classifier in conjunction with two feature selection methods for an overall accuracy of up to 0.534, which represents a large improvement over the majority class baseline of 0.188.

We also demonstrated how the user metadata can be incorporated in our classifiers. Although concatenation of features drawn both from blog content and profile elements did not yield any clear improvements over the best individual classifiers, we found that stacking improves the prediction accuracy to an overall accuracy of 0.643, as measured on our test dataset. A more in-depth analysis showed that not all industries are equally easy to predict: while industries such as *Real Estate* and *Religion* are clearly distinguishable with accuracy figures over 0.80, others such as *Banking* are much harder to predict.

Finally, we presented a qualitative analysis to provide some insights into the language of different industries, which highlighted differences in the top-ranked words in each industry, word semantic similarities, and the relative frequency of emotionally charged words.

## 7. ACKNOWLEDGMENTS

This material is based in part upon work supported by the National Science Foundation (#1344257) and by the John Templeton Foundation (#48503). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the John Templeton Foundation.

## 8. REFERENCES

- [1] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.
- [2] D. Bamman, C. Dyer, and N. A. Smith. Distributed representations of geographically situated language. 2014.
- [3] K. A. Bender, S. M. Donohue, and J. S. Heywood. Job satisfaction and gender segregation. *Oxford economic papers*, 57(3):479–496, 2005.
- [4] B. Bernstein. Language and social class. *The British journal of sociology*, 11(3):271–276, 1960.
- [5] B. B. Bernstein. *Class, codes and control: Applied studies towards a sociology of language*, volume 2. Psychology Press, 2003.
- [6] K. Bharat, S. Lawrence, and M. Sahami. Generating user information for use in targeted advertising, Jan. 12 2016. US Patent 9,235,849.
- [7] F. Celli. Unsupervised personality recognition for social network sites. In *Proc. of Sixth International Conference on Digital Society*, 2012.
- [8] G. Coppersmith, M. Dredze, and C. Harman. Quantifying mental health signals in twitter. *ACL 2014*, page 51, 2014.
- [9] J. Fink and A. Kobsa. A review and analysis of commercial user modeling servers for personalization on the world wide web. *User Modeling and User-Adapted Interaction*, 10(2-3):209–249, 2000.
- [10] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.
- [11] W. L. French. Can a man’s occupation be predicted? *Journal of Counseling Psychology*, 6(2):95, 1959.
- [12] G. Giacinto and F. Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19(9):699–707, 2001.
- [13] G. B. Gil, A. B. de Jesús, and J. M. M. Lopéz. Combining machine learning techniques and natural language processing to infer emotions using spanish twitter corpus. In *Highlights on Practical Applications of Agents and Multi-Agent Systems*, pages 149–157. Springer, 2013.
- [14] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12, 2009.
- [15] B. Guibert, J. Laganier, and M. Volle. An essay on industrial classifications. *Economie et statistique*, 20:1–18, 1971.
- [16] B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, pages 451–500, 2014.
- [17] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber’s heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM, 2011.
- [18] M. Hernandez, K. Hildrum, P. Jain, R. Wagle, B. Alexe, R. Krishnamurthy, I. R. Stanoi, and C. Venkatramani. Constructing consumer profiles from social media data. In *Big Data, 2013 IEEE International Conference on*, pages 710–716. IEEE, 2013.
- [19] Y. Huang, L. Yu, X. Wang, and B. Cui. A multi-source integration framework for user occupation inference in social media systems. *World Wide Web*, 18(5):1247–1267, 2015.
- [20] J. P. Johnson. Targeted advertising and advertising avoidance.

*The RAND Journal of Economics*, 44(1):128–144, 2013.

- [21] R. Kashyap and A. Nahapetian. Tweet analysis for user health monitoring. In *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*, pages 348–351. IEEE, 2014.
- [22] A. Kokkos and T. Tzouramanis. A robust gender inference model for online social networks and its application to linkedin and twitter. *First Monday*, 19(9), 2014.
- [23] W. Labov. *Sociolinguistic patterns*. Number 4. University of Pennsylvania Press, 1972.
- [24] W. Labov. *The social stratification of English in New York city*. Cambridge University Press, 2006.
- [25] V. Lamps, D. Preotiuc-Pietro, and T. Cohn. A user-centric model of voting intention from social media. In *ACL (1)*, pages 993–1003, 2013.
- [26] J. Li, A. Ritter, C. Cardie, and E. H. Hovy. Major life event extraction from twitter based on congratulations/condolences speech acts. In *EMNLP*, pages 1997–2007, 2014.
- [27] J. Li, A. Ritter, and E. H. Hovy. Weakly supervised user profile extraction from twitter. In *ACL (1)*, pages 165–174, 2014.
- [28] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [29] E. Mohammady and A. Culotta. Using county demographics to infer attributes of twitter users. *ACL 2014*, page 7, 2014.
- [30] J. Oberlander and S. Nowson. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 627–634. Association for Computational Linguistics, 2006.
- [31] D. Preotiuc-Pietro, V. Lamps, and N. Aletras. An analysis of the user occupational class through twitter content. The Association for Computational Linguistics, 2015.
- [32] D. Rao, M. J. Paul, C. Fink, D. Yarowsky, T. Oates, and G. Coppersmith. Hierarchical bayesian models for latent attribute detection in social media. *ICWSM*, 11:598–601, 2011.
- [33] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.
- [34] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.
- [35] P. Schmidt and R. P. Strauss. The prediction of occupation using multiple logit models. *International Economic Review*, pages 471–486, 1975.
- [36] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.
- [37] S. Volkova, G. Coppersmith, and B. Van Durme. Inferring user political preferences from streaming communications. In *ACL (1)*, pages 186–196, 2014.
- [38] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

## APPENDIX

### A. ADDITIONAL EXAMPLES OF WORD SIMILARITIES

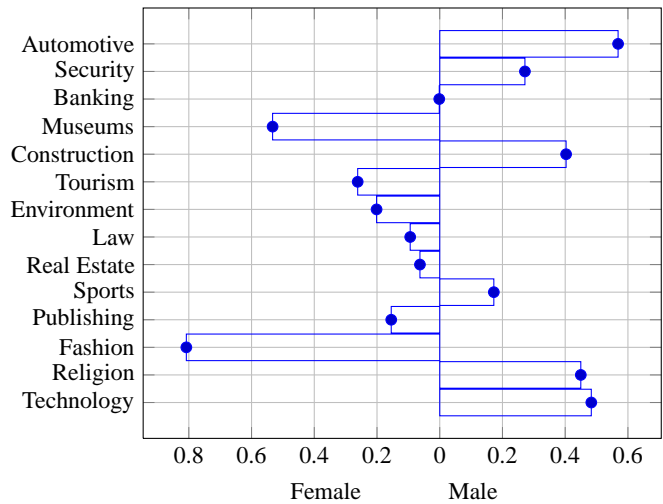
Religion		Sports	
term	cosine	term	cosine
professional	1.000	professional	1.000
mentoring	0.774	sports	0.833
education	0.745	coaching	0.801
niche	0.724	active	0.795
conversational	0.722	competitive	0.793
vocational	0.721	becoming	0.789
learner	0.720	major	0.785
educational	0.714	fellow	0.778
lock-ins	0.714	having	0.775
thorough	0.713	coaches	0.768

**Table 11: Terms with the highest cosine similarity to the term professional.**

Technology		Fashion	
term	cosine	term	cosine
leisure	1.000	leisure	1.000
playrooms	0.651	presale	0.752
photo-editing	0.650	versona	0.750
multi-media	0.647	jewelry	0.748
match-making	0.646	high-end	0.748
pre-ordered	0.644	sketchers	0.747
tradeshows	0.643	craft	0.743
tfp	0.643	vintage-inspired	0.738
schmooze	0.641	spruill	0.737
upload/download	0.640	baggu	0.733

**Table 12: Terms with the highest cosine similarity to the term leisure.**

### B. GENDER DOMINANCE IN INDUSTRIES



**Figure 3: Gender dominance for the different industries.**