Machine Learning in Federal Hiring:

How natural language processing and transfer learning can improve equity and efficiency in government jobs

Introduction

Overview of the Federal Hiring Website

The US Federal government (USG) employs approximately 2,100,753 workers (Jennings and Nagel, 2019) and any of these 2 million who began their work after 2010 were hired through a single website: USAjobs.gov. This portal hosts tens of thousands of jobs on any given day, each post detailing things like responsibilities, qualifications, and grade. These jobs have been managed almost entirely manually, requiring HR employees in each department to set their own job requirements.

General Service Job Grade Assignment

While some hiring tasks are highly complex, such as defining the specific duties of a job, others are more amenable to automatization. One task is both especially cumbersome to humans and quite manageable for machines: the assignment of job grade. The USG's Office of Personnel Management (OPM) uses several job grade categories, but for the purposes of this paper, we will focus on the most common: General Schedule (GS). General Schedule jobs are graded from GS-1 to GS-15 based on responsibility and seniority, with pay bound to the GS level and adjustments allowed for location and years spent in the position (U.S. Office of Personnel Management, 2019).

Each job is assigned a specific job code derived from OPM's guiding documents[1] (U.S. Office of Personnel Management, 2009; U.S. Office of Personnel Management, 1991; U.S. Office of Personnel Management, 1998). These documents are used by agencies to classify the job into the correct group and specific series.

0300 – General, Administrative, Clerical, and Office Services Group

Miscellaneous Administration and Program Series**	0301	Support Services Administration Series*	0342
Messenger Series*	0302	Management and Program Analysis Series**	0343
Miscellaneous Clerk and Assistant Series**	0303	Management and Program Clerical and Assistance Series*	0344
Information Receptionist Series*	0304	Logistics Management Series**	0346
Mail and File Series*	0305	Equipment Operator Series*	0350
Government Information Series**	0306	Data Transcriber Series*	0356
Records and Information Management Series**	0308	Equal Opportunity Compliance Series*	0360
Correspondence Clerk Series*	0309	Equal Opportunity Assistance Series*	0361
Work Unit Supervising Series**	0313	Telephone Operating Series*	0382
Secretary Series*	0318	Telecommunications Processing Series*	0390
Closed Microphone Reporting Series**	0319	Telecommunications Series*	0391
Clerk-Typist Series**	0322	General Telecommunications Series**	0392
Office Automation Clerical and Assistance Series**	0326	Communications Clerical Series**	0394
Computer Operation Series*	0332	Administration and Office Support Student Trainee Series	0399
Computer Clerk and Assistant Series*	0335		
Program Management Series	0340		
Administrative Officer Series**	0341		

Figure 1. Excerpt from the *Handbook of Occupational Groups and Families*

For example, the 0300 job group encompasses "General Administrative, Clerical, and Office Services". Within the 0300 group, the specific job codes could include "Data Transcriber, Telecommunications processing" and "Secretary" among others. Each of these series then have more information related to what

[1] GS level assignment is prescriptively discussed in the following documents: Position Classification Standards, Classifier's Handbook and the GS Supervisory Guide

is specifically included in that job description. An excerpt from the Handbook of Occupational Groups and Families can be seen in Figure 1.

The implications of poor labeling

The standard application of grades across all agencies and job types, therefore, is crucial to the government's policy of equity. If two jobs have equivalent responsibilities, they are required to be the same GS-level regardless of what Agency the job is in. These are found in three government documents: Introduction to the Position Classification Standards (U.S. Office of Personnel Management, 2009), General Schedule Supervisory Guide (U.S. Office of Personnel Management, 1998), and The Classifier's Handbook (U.S. Office of Personnel Management, 1991). Consistent classification is difficult to achieve through hundreds of different human resource specialists guided by three tedious technical documents full of caveats and conditions.

Potential benefits of this research project

Artificial intelligence has long been used to classify texts, and in this case, it can do the same to standardize job grade assignment in USG hiring. This paper explains the methodology and application of a successful experiment to prescribe job grade of a General Service job using natural language processing techniques (NLP) to analyze the content of the job post. The successful application of this approach would mean faster, more efficient, and more equitable management of job postings in the United States federal government.

Structure of the paper

This paper explores related application of similar machine learning methods, the existing and accessible data from USAjobs.gov, and the methodology used to train the final model. Finally, we assess the model's results and implications and suggest future work that could improve upon these findings.

Background and Related Work

Document Classification Using NLP

Natural language processing has long been used to classify documents based on the text content. In 2001, Steven Bird and Edward Loper released the first version of NLTK (natural language toolkit), a common Python library used in NLP even today (Bird, 2017). In 2009, Ramdass and Seshasai used its Naive Bayes and Maximum Entropy classification to classify news articles by category with .77 and .72 accuracy, respectively (Ramdass & Seshasai, 2009). Ram and Prasanna used Neural Networks (NNs) to classify text (Ram & Prasanna, 2013), emphasizing that mass amounts of training data can improve accuracy, a finding echoed by Huang and Chen in their text classification work with Deep NNs. In particular, text vectorization is a common technique to break down words and sentences into a format more easily fed into machine learning models such models as NNs (Krishan & Kamath, 2018).

Most recently NLP has been used to explore even more complex phenomena, such as crowd mentality in the stock market (Feuerriegel & Gordon, 2018) and emotion recognition (Kratzwald et al., 2018).

Resume Classification Using NLP

NLP techniques have also been applied to the hiring process, but most frequently to classify resumes. Yu, Guan and Zhou highlighted the use of stacked models to accomplish this task (Yu et al., 2005). Sayfullina, Malmi, Liao and Jung trained Convolutional NNs on job description snippets to classify resume data and tested the results on labeled resume data (Sayfullina et al., 2017).

Natural Language Applications in the Government

In the US Government, NLP has begun to be applied in some departments, such as the Department of Defense, to assist analysts at DARPA (Eggers et al, 2019; Onyshkevych, n.d.), and the National Institutes of Health, to identify public health behaviors (Afshar et al, 2019). It is further used by external analysts to advise government on complex tasks like managing policy suggestions (Hagen et al, 2015). Unfortunately,

outside of defense and medical research, publicly known instances of NLP applications in government are not common.

Computational methods

Word Embeddings

World embedding, or the vectorization of words, is commonly used to "map" words in a multidimensional conceptual space. TensorFlow, a widely used ML library, explains that word embedding models "represent (embed) words in a continuous vector space where semantically similar words are mapped to nearby points ('are embedded nearby each other')" (TensorFlow, n.d.). They are based on the Distributional Hypothesis, which assumes that when words continually appear in the same contexts, they must share similar semantic meanings (Mikolov et al, 2013).

Transfer learning

Word embedding packages across machine learning platforms commonly come pre-trained on other language datasets, which significantly cuts down on the time the model needs to train (Google, 2013; H20.ai, 2019). This technique is called transfer learning, wherein a new model is able to build on the the calibrations that resulted from a previous model's training on separate data. It saves time and can improve the robustness of a model whose dataset is less than millions of observations. In NLP, we can define transfer learning in layman's terms as using an algorithm that already "speaks English", and only needs to "learn" the specifications of the task at hand, as opposed to having to teach a model English from zero.

Maximum Likelihood Regression Tree

Least squares (LS) regression trees minimize square loss at each branch to achieve minimal risk and eventually the lowest mean squared error for the regression at hand (Breiman, 1984). Trees are often "overgrown", or overfit, and then pruned backwards based on the ratio between the error and the size of the tree (the objective).

A maximum likelihood regression tree (MLRT) inherits this general idea, but incorporate model selection criteria, likelihood ratio tests, and other likelihood-based methods into each recursive partition. According to Su et al., who proposed the method in 2004, MLRT improve upon LS regression trees in several ways: "Compared with other least squared tree methods, maximum likelihood regression trees (MLRT) reject the use of many ad hoc approaches and rely on more established methods; they have easy extension to handle data involving other types of responses; in addition, simulation study shows that MLRT tends to provide more accurate tree size selection than CART" (Su et al, 2004).

Data exploration

Data from USAjobs.gov

Each job posted on USAjobs.gov includes information in a standardized format with at least the following sections:

- Position title and ID
- Department information
- Position start, end, posting and closure dates
- User Area Details
- Qualification summary
- Location

Crucially, "User Area Details" includes the high and low range of the job grade (shown as "High Grade" and "Low Grade" respectively), as well as number of openings, travel requirements, and job summary. Historical records are also available at the Developer page of USAJobs.gov after submitting a request and being

granted an API key (USAjobs.gov, n.d.), however not all fields are available. To test the viability of NLP in predicting the grade of a job based on the full text content of the job post, we scraped the full job postings for all openings active on February 26th, 2019.

Data cleaning

The data was pulled from the USAJobs.gov website, which exposed multiple endpoints for scraping data. The endpoint of interest to us was the search endpoints, or the job posts that could be returned in a job search. These endpoints returned json formatted payloads for each job listing. Using the search API, we retrieved nearly 14,000 current jobs available on USAJobs.gov.

Converting the data from JSON to CSV created additional due to the one-to-many relationships contained in numerous fields. For example, a single job could reference multiple locations. After scraping these records, we converted them into multiple normalized database tables, which allowed us to easily query the necessary data for future modeling.

Despite the data being easily available through a website API, there were significant irregularities in the data that made data analysis difficult. There were numerous missing values, hidden characters, and poor data entry that significantly affected the ability to properly store the data into a single data frame. The data was finally able to be sorted into data tables using a hidden character to separate records, and dropping rows that contained insurmountable errors such as missing key text fields.

Lastly, before feeding the textual data into the machine learning algorithms, the text needed to be normalized in a standard manner. This involves trimming white space, stemming words, removing special characters, removing stop words, and, in our case, removing numerical values. Stop words include common words ('GS', 'and', 'they', etc.) whose occurrence was frequent enough but insignificant enough to interfere with the performance of the model. Removing numerical values was critical to prevent data leakage, as a common indicator of job level is the number of years spent at the previous level lower level. For example, a GS-14 requires at least one year at a GS-13 level. Eliminating numerical data from our training data prevented this from being an issue.

These data cleaning steps allow the feature engineering algorithms to extract the most relevant features from the text and prevents overfitting due to feature leakage of numerics directly related to the target variable.

The Final Dataset

The final dataset of 13,955 posts was curated from all available current posts, occupying 84MB of data. It included 3,955 unique job postings, 11,312 unique qualification summaries across 25 government departments and 327 organizations.

The table below shows an overview of job postings and grade information by department.

Department Name	Number Postings	Avg. low grade	Avg. high grade	Avg. grade range
Department of the Air Force	2,075	6.55	10.38	3.83
Department of Transportation	160	9.33	10.69	1.44
Department of the Treasury	177	9.34	10.61	1.27
Department of Health and Human Services	736	9.52	10.65	1.14
Department of Labor	69	8.17	9.28	1.1
Department of Commerce	539	4.29	5.39	1.09
Executive Office of the President	11	11.64	12.64	1
Department of Energy	75	6.05	6.99	0.93
Department of Agriculture	84	9.29	10.12	0.83
General Services Administration	63	10.68	11.51	0.83

Department of the Navy	1,026	6.36	7.14	0.77
Department of Justice	204	9.22	9.94	0.72
Other Agencies and Independent Organizations	518	9.37	10.01	0.64
Judicial Branch	33	6.27	6.88	0.61
National Aeronautics and Space Administration	49	11.49	12.06	0.57
Department of the Interior	570	7.54	8.1	0.56
Department of Housing and Urban Development	23	10.39	10.87	0.48
Department of Homeland Security	536	10.48	10.94	0.46
Department of the Army	2,940	8.09	8.45	0.35
Department of Defense	757	8.79	9.06	0.27
Department of Veterans Affairs	3,203	6.59	6.83	0.24
Legislative Branch	84	7.18	7.35	0.17
Department of State	19	10.22	10.33	0.11
Court Services and Offender Supervision Agency for DC	3	13	13	0
Department of Education	1	9	9	0

Table 1. GS Grade Information by Department

Importantly, Table 1 shows which USG Departments have the largest range between the maximum and minimum job grade offered for a given position. We will refer to this measure as the average grade range. Again, the Navy and the Air Force lead in this regard

Modeling methodology

Software Packages

We used a wide variety of software to develop the model. Much of the EDA (Exploratory Data Analysis) and data preparation was done using python and tableau. Eventually, ML.Net, a Microsoft owned machine learning package written in C#, was chosen as the primary tool for the final model's entire pipeline. This is because the package has sourced many of the best open source advancements and placed them in easy-to-use API's that can easily be ported to a production environment.

Model Set-up

As each job posting includes a high and low job grade, two models were built; one to predict high grade, and another to predict low. It is worth noting that in a significant majority of postings – 71% – high and low grade have the same value; however, in the remaining 29%, the average range between high and low values is 3.42. For this reason, we would have to train two models: one with the target variable being low grade and one with the target variable being high grade.

The cleaned and normalized text in the field 'Qualification Summary' is a simple and standardized input that would make a powerful input for the prediction task at hand. As explained in section 3.2, we ran the qualification summary text through a standard text normalization algorithm that removes numerals, special characters, and extra white space. This method has one drawback: the normalization removes some useful information in the text, such as number of years' experience required: if a posting includes the sentence "This job requires 5 years of experience," the most meaningful information would be eliminated and the phrase would have no meaning. It would be better to use more advanced data cleaning methods to determine which numerals cause data leakage and which do not; however, due to the completely free format of the training text, this was prohibitively difficult, and so we preferred a more cautious method of removing all numerals.

GloVe Vectorization Technique

A transfer learning word embedding model is a perfect fit to classify job grade based on text within the job posting. In the case of the USAjobs.gov dataset, it was especially useful as there are less than 14,000 observations, and a pre-trained model would improve accuracy significantly. Stanford's GloVe (Global Vectors for Word Representation) provided an unsupervised vectorization algorithm pre-trained on two billion tweets[2] (Pennington et al, 2014). In this algorithm, cosine similarity between words is used to measure semantic similarity, and nearest neighbors are evaluated to engineer new features whose values quantify the relatedness of words.

Decision Tree Ensemble

Once the vector features have been engineered, they were used to train a Microsoft FastTree regression model (Microsoft, n.d.), a maximum likelihood decision tree, to predict the job grade. The model predicts a number between 1 and 15, but not necessarily a whole number, to allow for more flexibility in the predictions. As explained above, one model was trained to predict "low grade" and another to predict "high grade".

Given the relative simplicity of the qualification summary text, this model is very powerful – .92 accuracy – and able to prescribe job grade in a more standard way than a human, bogged down in details, would be able to do.

Model Pipeline

Below is the entire machine learning pipeline:

Figure 2. Model Pipeline in Summary

To apply the model to a new job posting, the text of a raw Qualification Summary would be fed into the pipeline. It would subsequently be normalized, tokenized, cleaned of stop words, fit with word embeddings, and finally regressed using the fast tree algorithm.

Model results

Variance by Job Code

With different policies, internal cultures and types of work carried out across agencies and fields, it is logical that the model would have varying performance across jobs types and departments. To better understand this variance, we limited our analysis to jobs that had more than 10 occurrences in the test data set. Among those, the high and low performers can be found in the tables below.

Table 2. Best Performers by Job Code

Job Name	Number jobs	Avg. abs. error: Low grade	Avg. abs. error: High grade	Avg. low grade	Avg. high grade
Food inspection	12	2.36	0.36	5.00	7.00

[2] The algorithm provides multiple options for pre-training corpus, but for this model we selected the 50d Twitter corpus. See https://nlp.stanford.edu/projects/glove/ for more information.

Industrial Engineering	131	0.83	0.45	10.90	11.45
Toxicology	12	0.46	0.51	13.92	14.00
Mathematical Statistics	16	3.45	0.51	9.00	12.06
Aerospace Engineering	121	0.97	0.55	10.78	11.45

Table 2. Best Performers by Job Code

Table 3. Worst Performers by Job Code					
Job Name	Number jobs	Avg. abs. error: Low grade	Avg. abs. error: High grade	Avg. low grade	Avg. high grade
Management and Program Clerical and Assistance	15	2.68	2.27	5.93	7.47
Program Management	29	2.27	2.11	13.59	13.93
Budget Clerical and Assistance	16	2.67	1.98	5.50	7.38
Social Science Aid and Technician	21	2.70	1.83	4.86	7.10
Dietitian and Nutritionist	38	2.07	1.83	8.92	10.26

Table 3. Worst Performers by Job Grade

The job codes whose high and low grade that the model was most accurately able to predict are not those on the high or low end of the GS scale. In fact, the top three job categories on which the model performed best have average job grades of 6, 11, and 14 – a wide range.

However, there does appear to be a differentiation in the types of skills required. For example, "Industrial Engineering", "Mathematical Statistics" and "Aerospace Engineering" are have key skillsets that are relatively concrete and easy to quantify – "hard skills" as they are sometimes called. For Example, 2162 job postings, including every Aerospace Engineering job in the dataset, contain the word "thermodynamics". 66% of these jobs have a high grade of either GS-12 or GS-13, meaning that even the presence of this one word significantly improves the predictability of the high grade.

Conversely, jobs such as "Management and Program Clerical and Assistance" or "Program Management" include requirements that are less tangible and described in words that are not unique to that level of seniority. The word "management" can be used in phrases like "management of daily reporting," which is probably a low-seniority task, and in phrases like "direct management of all personnel," a very senior task. The distribution of the keyword "management" also has a longer tail and is left-skewed, reducing the predictive power of this word. More likely, these jobs require a certain number of years of experience doing these types of tasks. Here we find a case that clearly demonstrates the problem of removing all numerical values to avoid data leakage.

This pattern was repeated across multiple technical keywords such as "statistics", "engineering", "probability", and less technical words such as "administrative", "scheduling", and "administrative" it becomes easy to understand why technical jobs have a more predictable GS level. A comparison of "Processing" and "Thermodynamics" can be seen in the figure below.

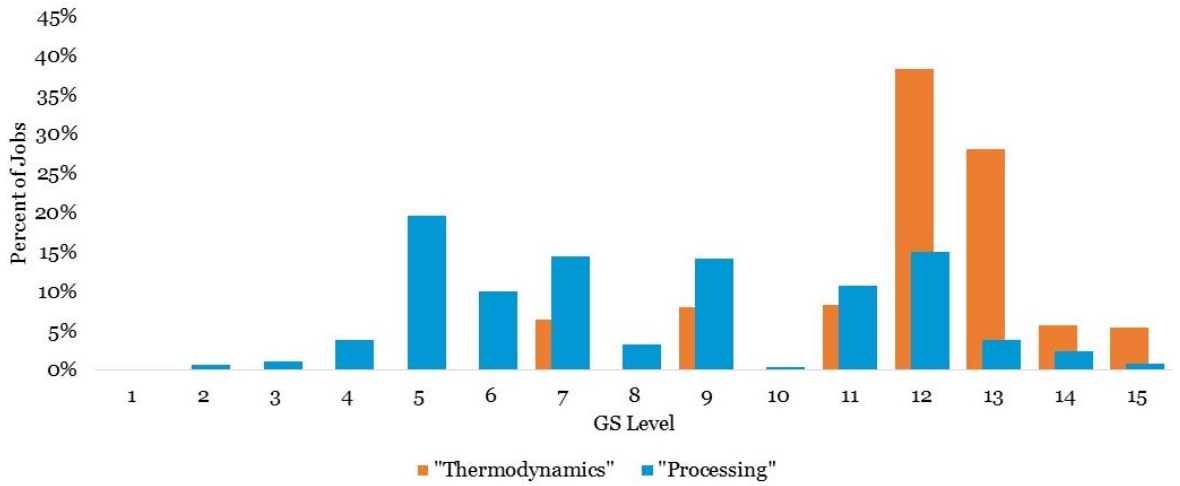

Figure 3. Distribution of select keywords across GS levels

Variance by Agency

Another way to measure model performance is how it did across agencies. Ideally, every agency would have relatively similar average error rates. However, given that there are different individuals assigning job grades in each agency, this is unlikely.

In the table below we can see that the Department of the Treasury and Housing and Urban Development (HUD) have the best performance in the model, while the US Airforce and Department of Energy have the worst. This insight is less intuitive than those of the job series' themselves, so, to evaluate this variance further, the Air Force will be used as a case study in a "deep dive" to understand these results.

Table 4. Best Performers by Department					
Name	Number jobs	Avg. abs. error: Low grade	Avg. abs. error: High grade	Avg. low grade	Avg. high grade
Department of the Treasury	165	1.80	0.82	9.44	10.78
Department of Housing and Urban Development	19	1.34	0.83	12.58	13.16
General Services Administration	63	1.37	0.87	10.68	11.51
Department of Justice	145	1.54	0.94	9.37	10.24
Department of the Army	1814	1.29	1.01	9.79	10.21

Table 4. Best Performers by Department

Table 5. Worst Performers by Department					
Name	Number jobs	Avg. abs. error: Low grade	Avg. abs. error: High grade	Avg. low grade	Avg. high grade
Department of the Air Force	1149	2.74	1.90	7.85	10.82
Department of Energy	44	2.13	1.77	9.32	10.84
Executive Office of the President	11	1.38	1.71	11.64	12.64
Legislative Branch	39	1.62	1.65	8.92	8.95

National Aeronautics and Space Administration	47	1.49	1.48	11.98	12.57

Table 5. Worst Performers by Department

Air Force Case Study

Among the Air Force job postings, the top 15 worst performing individual jobs from the model are shown below:

Position Title	Low Grade	Prediction Low	High Grade	Prediction High
Clinical Nurse	5	7.78	15	7.78
Equipment Specialist	5	6.33	13	6.33
LEAD PROGRAM ANALYST	13	6.71	13	6.71
PROGRAM ANALYST	13	6.71	13	6.71
Explosives Safety – Direct Hire Authority	1	9.21	15	9.21
Safety and Occupational Health Management – Direct Hire Authority	1	9.21	15	9.21
Community Planning – Direct Hire Authority	1	9.21	15	9.21
Environmental Protection Specialist – Direct Hire Authority	1	9.21	15	9.21
Environmental Protection Assistant – Direct Hire Authority	1	9.21	15	9.21
Sports Specialist – Direct Hire Authority	1	9.21	15	9.21
Chaplin – Direct Hire Authority	1	9.21	15	9.21
Security Administration – Direct Hire Authority	1	9.21	15	9.21
Fire Protection and Prevention – Direct Hiring Authority	1	9.21	15	9.21
Police – Direct Hire Authority	1	9.21	15	9.21
Security Guard – Direct Hire Authority	1	9.21	15	9.21

Table 6. Worst Performing Jobs within the Air Force

The first four jobs clearly show poor predictive power, indicating that Clinical Nurses, Equipment Specialists and Program Analysts are difficult to predict. This may have to do with the fact that the language used to describe these jobs' qualifications and duties does not change significantly with increased seniority – here again we run into the problem of missing years of experience due to eliminated numerical values. A lead nurse with 15 years of experience will be described with similar vocabulary as a nurse beginning his first job.

The remaining 11 jobs are assigned job grade ranges from 1 to 15, meaning that they encompass the entire pay scale. In terms of test data, this provides no meaningful information, and is effectively a null value. To explain why, it is worth looking at the titles of the postings in question - while these jobs range from "Environment Protection Assistant", to "Police", to "Explosives Safety", they are all designated "Direct Hire Authority".

Direct Hire Authority (DHA) is, according to OPM, "an appointing (hiring) authority that OPM can give to Federal agencies for filling vacancies when a critical hiring need or severe shortage of candidates exists" (U.S. Office of Personnel Management, n.d.). This means that jobs with DHA granted could simply hire someone at any level to fill the position. If there is an immediate need for data entry clerks, a typically low paying job, DHA could be granted and the clerks could be hired at GS-15 levels. Obviously this free-reigning authority would be difficult to predict as it has no relevance to the qualifications of the jobs itself.

Direct Hiring Authority

One reason that the Air Force may be performing poorly is if it has a high number of DHA jobs exist in the Air Force. Below is a table of all DHA jobs in the dataset:

Agency	Number of DHA Jobs
Department of the Air Force	147
Department of Transportation	7
Department of Health and Human Services	7
Department of Veterans Affairs	4
Department of Homeland Security	3
Department of the Interior	3
Department of the Army	3
Department of Commerce	2
Department of Energy	1
Department of the Treasury	1

Table 7. DHA Jobs by Department

Clearly, the Air Force is the biggest poster of DHA jobs - as the owner over 80% of all DHA opportunities in the federal government. Once the DHA jobs are removed from the dataset, the Average error rate for Air Force jobs drops significantly - from 1.90 to 1.3. This drops the Air Force's rank from being the worst, to middle of the pack (7th of 22).

Air Force Error Rate		
	Average Error- High Grade	Rank
All Jobs	1.90	1st (worst)
DHA Only	5.79	1st
DHA Removed	1.31	7th

Table 8. Air Force Error Ranking

Other Problems with the Training Data

Other data quality issues exist outside of the potential DHA issue above. 171 jobs had a low grade or high grade listed as "zero". Zero is not an allowable grade on the GS 1 - GS 15 pay scale. Since the model was trained on this incorrect data, it likely affected its performance as the model had to account for "GS-0" jobs.

Additionally, 154 jobs had a low GS level of 1 and a high GS level of 15. There is no job in the federal government that should have such a wide range of potential pay grades. This is likely due to user error in inputting an allowed value despite it being incorrect.

These data quality issues need to be addressed in data-entry within OPM to improve the quality of scoring before implementing automated job grade assignment.

Conclusions and Future Work

Opportunities for Improvement

In spite of obstacles stemming from inconsistent training data, this methodology has proved quite accurate in predicting job grade in the GS system of the federal government. Nonetheless, it is possible to improve upon the model by incorporating more varied inputs, such as job description and job title, in an ensemble model. Within the sub-models of such an ensemble, other features could be engineered beyond the original word embeddings, such as more domain-specific categories that could be tailored to the USAjobs.gov website with the knowledge of USG human resources expert.

The model was trained using job postings available on a single day; more robust historical data could decrease variance within departments or job codes, and either support or disprove current hypotheses on mismanagement of GS classification to date.

Finally, the model could be refined with insight from an industry expert. It is likely that understanding of US Federal Government hiring and departments could inform a more tailored model, just as it could inform the engineering of domain-specific features. We offer this model as a base to be refined and built upon by those with domain expertise and access to historical USAjobs.gov job posting records.

Implications for practice

As machine learning techniques gain traction among early adopters, most government entities lag behind, even as these techniques would be highly beneficial to their practice. While certain departments within the US federal government are more research oriented, such as the Department of Defense, others, such as OPM, are spending valuable time and resources on tasks that could be automated to increase efficiency, accuracy and standardization. OPM should evaluate both the implications this model has in exposing inequity within their systems as well as the potential to boost productivity in human resource management.

Implications for theory

Even though the hypothetical applications of machine learning and NLP techniques in government have been explored for some years now, only a few experimental designs have been put to the test, most frequently at a local level or a very limited scope (Leonard, 2018). The successful application of this model to a nationwide federal government problem proves that machine learning not only should be implemented to solve problems in the government sector, but that it can be done with strong results and relative ease. This paper adds to the scattered literature that supports NLP techniques in government with a clear, concrete example of a highly successful model ready for near-immediate implementation.

References

Afshar, M., Phillips, A., Karnik, N., Mueller, J., To, D, Gonzalez, R et al. 2019. "Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation," *Journal of the American Medical Informatics Association* (26:3).

Bird, S. 2017. "nltk/nltk/wiki/FAQ" on *GitHub*. Retrieved from https://github.com/nltk/nltk/wiki/FAQ

Breiman, L., Friedman, J., Stone, C., Olshen, R. 1984. *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall/CRC.

Eggers, W., Malik, N., Gracie, M. 2019. "Using AI to unleash the power of unstructured government data: Applications and examples of natural language processing (NLP) across government". Retrieved from https://www2.deloitte.com/insights/us/en/focus/cognitive-technologies/natural-language-processing-examples-in-government-data.html

Feuerriegel, S. & Gordon, J. 2018. "Long-term stock index forecasting based on text mining of regulatory disclosures," *Decision Support Systems* (112), pp. 88-97.

Google. 2013. word2vec. Retrieved April 23, 2019 from https://code.google.com/archive/p/word2vec/

H2o.ai. 2019. Word2vec. Retrieved April 23, 2019 from http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/word2vec.html

Hagen, L., Uzuner, O., Kotfila, C., Harrison, T., Lamanna, D. 2015. "Understanding Citizens' Direct Policy Suggestions to the Federal Government: A Natural Language Processing and Topic Modeling Approach," in *48th Hawaii International Conference on System Sciences*, pp. 2134-2143.

Huang, Y. & Chen Blind, J. 2013. "Classifying the text using the power of deep learning," *International Journal of Engineering and Technology*, pp. 3153-3156.

Kratzwald, B., Ilic, S., Kraus, M., Feuerriegel, S., & Prendinger, H. 2018. "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decision Support Systems* (115), pp. 24-35.

Krishan, G. & Kamath, S. 2018. "A Supervised Learning Approach for ICU Mortality Prediction Based on Unstructured Electrocardiogram Text Reports," in *Natural Language Processing and Information Systems*, Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., Meziane, F. (eds). NLDB 2018. Lecture Notes in Computer Science (10859). Springer, Cham.

Leonard, M. 2018. "Government leans into machine learning," *GCN*. Retrieved from https://gcn.com/pages/about.aspx

Microsoft. n.d. "FastTreeRegressionTrainer Class". Retrieved from https://docs.microsoft.com/en-us/dotnet/api/microsoft.ml.trainers.fasttree.fasttreeregressiontrainer?view=ml-dotnet

Mikolov, T., Chen, K., Corrado, G., Dean, J. 2013. "Efficient Estimation of Word Representations in Vector Space". Retrieved April 22, 2019 from https://arxiv.org/pdf/1301.3781.pdf

Onyshkevych, B. n.d. "Deep Exploration and Filtering of Text (DEFT)". Retrieved April 27, 2019 from https://www.darpa.mil/program/deep-exploration-and-filtering-of-text

Pennington, J., Socher, R., Manning, C. 2014. "GloVe: Global Vectors for Word Representation". Retrieved from https://nlp.stanford.edu/projects/glove/

Ram, V. & Prasanna, S. 2013. "A unique way of measuring the similarity of the documents using neural networks," *International Journal of Engineering Research and Development* (2:6), pp. 397-401.

Ramdass, D. & Seshasai, S. 2009. "Document Classification for Newspaper Articles". Retrieved from https://pdfs.semanticscholar.org/aa96/9114cf6e4d77c5bb3dd62a20bee3446f33ab.pdf

Sayfullina, L., Malmi, E., Liao, Y., Jung, A. 2017. "Domain Adaptation for Resume Classification Using Convolutional Neural Networks" in *Analysis of Images, Social Networks and Texts*, van der Aalst, W. et al. (eds). AIST 2017. Lecture Notes in Computer Science, (10716). Springer, Cham

Su, X., Wang, M., Fan, J. 2004. "Maximum Likelihood Regression Trees," *Journal of Computational and Graphical Statistics* (13:3), pp. 586-598.

TensorFlow. n.d. *Vector Representations of Words*. Retrieved April 28, 2019 from https://www.tensorflow.org/tutorials/representation/word2vec

U.S. Office of Personnel Management. 2019. *Salaries & Wages*. Retrieved from https://www.opm.gov/policy-data-oversight/pay-leave/salaries-wages/

U.S. Office of Personnel Management. 2009. *Introduction to the Position Classification Standards*. Retrieved from https://www.opm.gov/policy-data-oversight/classification-qualifications/classifying-general-schedule-positions/positionclassificationintro.pdf

U.S. Office of Personnel Management. 1998. *General Schedule Supervisory Guide*. Retrieved from https://www.opm.gov/policy-data-oversight/classification-qualifications/classifying-general-schedule-positions/functional-guides/gssg.pdf

U.S. Office of Personnel Management. 1991. *Classifier's Handbook*. Retrieved from https://www.opm.gov/policy-data-oversight/classification-qualifications/classifying-general-schedule-positions/classifierhandbook.pdf

U.S. Office of Personnel Management. (n.d.). *Hiring Information: Direct Hire Authority*. Retrieved April 3, 2019 from https://www.opm.gov/policy-data-oversight/hiring-information/direct-hire-authority/#url=Fact-Sheet

USAjobs.gov. n.d. *General – Overview*. Retrieved from February 24, 2019 from https://developer.usajobs.gov/General

Yu, K., Guan, G., Zhou, M. 2005. "Resume Information Extraction with Cascaded Hybrid Model" in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 499–506.