An SVM-Based Approach to Predicting Level of Job Anxiety in Corporate Professionals using Linguistic Markers on Twitter

Unnathi Utpal Kumar

Abstract Work anxiety is linked with decreased job commitment and satisfaction. Yet, work anxiety, like other mental health problems, is not physically diagnosable. The lack of diagnosis and cure of job anxiety leads to lower levels of economic productivity and adds to the mental health epidemic. This study proposes a machine learning model to predict a corporate professional's level of work anxiety using their tweets by identifying linguistic markers associated with work anxiety. The Twitter API was used to create a dataset of over 15,000 corporate professionals. Thousands of tweets were collected from these users over 3 periods of time (May–August 2019, May–August 2020, and January-April 2021). After conducting sentiment and linguistic analysis, tweets from 90 random users (manually labelled for their job anxiety scores according to the job anxiety scale) were used to train/test an SVM regression model. The model achieved an RMSE of 0.2 and an accuracy of 83%. This approach has the potential to enable early detection of work anxiety and alert individuals about their mental health.

Keywords Job anxiety · Social networks · Twitter · Natural language processing

1 Introduction

Job anxiety is a stimulus-related type of anxiety, occurring in the workplace or when thinking of the workplace [8]. Whilst the workplace can provide support [11], it can provoke anxiety too. High standards and expectations can lead to feelings of insufficiency. Social anxiety might be heightened due to monitoring and sanctioning, whilst competition with colleagues might lead to persecution fears [4]. With leadership, greater responsibilities can lead to overburdening, provoking anxiety [5].

Work anxiety is correlated with lower job commitment [1] can cause absenteeism and impact performance [7], further associated with minimized efficiency and increased expenses for businesses, insurances, and public pension funds.

U. U. Kumar (✉)
Pathways World School Aravali, Gurugram, India

J. K. Mandal et al. (eds.), *Topical Drifts in Intelligent Computing*, Lecture Notes in Networks and Systems 426, https://doi.org/10.1007/978-981-19-0745-6_13

Symptoms of work anxiety include typical anxiety symptoms, like blushing, trembling, or palpitations [8]. At present, despite its clinical and economical importance, only 1 questionnaire is available to measure job anxiety, known as the job anxiety scale (JAS) [6]. However, JAS's validity is difficult to assess since it is a self-report scale, such that it consists of shortcomings due to acquiescence and social desirability [9]. Therefore, with no specific ICD codes for work anxieties, it can be difficult to characterize and thus diagnose them dynamically.

In context of these challenges, this study examines and develops a novel solution based on the previous metrics of job anxiety. The research here analyzes social media as a tool to express work-related concerns and thus to predict work anxiety. As people use Twitter to express their thoughts, social media can become a data source for job anxiety. By training an AI-based model on this data, work anxiety can effectively be detected.

This study aims to investigate the following question: can a mathematical model identify and quantify the relationship between a healthy corporate professional's use of language in comparison with that of a corporate professional suffering from job anxiety?

The hypothesis pursued is, if a healthy corporate professional's use of language is compared to that of a corporate professional suffering from job anxiety, then there exists a difference in terms of the linguistic markers, which can be identified by a mathematical model such that the use of these linguistic markers is further heightened in individuals suffering from higher levels of work anxiety.

The main contributions in this paper include

1. A Python script using the Twitter API and Selenium was developed to create a comprehensive dataset with tweets from over 15,000 Twitter users
2. Sentiment analysis of tweets in relation to levels of job anxiety (as expressed linguistically) over 3-month periods in 2019, 2020, and 2021
3. An SVM regression model to predict levels of job anxiety from tweets

The solution proposed in this study, the SVM regression model, combats the problems associated with other ways to measure levels of work anxiety, as it avoids response bias by not requiring a conscious undertaking like a survey and is also specific to job anxiety than anxiety generally. The solution thus has the potential to increase economic productivity and help combat mental health problems.

2 Background Literature

Generally, only some research has been conducted regarding work anxiety, as researchers often adapt different pre-existing instruments or tools. Nonetheless, one empirical way to measure perceived job anxiety is through the job anxiety scale (JAS) [6]. JAS is a self-rating scale to measure job anxiety, containing 5 main dimensions, (i) stimulus-related anxiety and avoidance behaviour, (ii) social anxieties, (iii) health-related anxieties, (iv) cognitions of insufficiency, and (v) job-related worries.

However, the extensivity of the JAS renders it uneconomical and inefficient and places job anxiety in a very specific, unmalleable context. Therefore, to consider job anxiety as a wider construct for this study, a shortened, empirically derived version of the JAS, JAS-15 [10], was considered. The JAS-15 retains the 5 theoretically meaningful dimensions.

Whilst no research has been conducted on language use and work anxiety specifically, the previous research helps identify possible correlations between linguistic structures and anxiety. For example, the use of first-person singular pronouns during negative memory recall is associated with individuals suffering from anxiety [2]. It was also found that the use of unigrams like "withdrawal" and "severe" on Twitter indicated anxiety [3].

Whilst the background literature is useful in defining the study's terms, they either do not consider job anxiety specifically or have not been implemented in a context relevant to social media. This lack of data suggests that further research is needed to determine linguistic markers of work anxiety, along with an efficient and accurate solution to realistically predict work anxiety levels.

3 Data Collection

3.1 Data Collection

A list of Twitter usernames of corporate professionals had to be collated.

First, a list of 10 keywords that corporate professionals might use in their tweets was created, identified based on prior knowledge.

Thereafter, using "search" on Twitter, it was manually checked that, for each keyword, most initial tweets from search results were by corporates. Then, using Web scraping with Python and Selenium (tool automating browsers), each keyword was searched and first 3000–3200 tweets collected. The no. of tweets could not be standardized due to memory overflow issues on Microsoft Edge (driver browser), but no less than 3000 and no more than 3200 tweets were collected from each keyword. The content, username, and name associated with tweets in each keyword were stored in CSVs.

However, not all usernames collected were those of corporate professionals, so irrelevant data had to be removed using filtering on Microsoft Excel. First, all duplicate usernames were removed and usernames containing the words "news" or "job" or links were removed (generally portals and not individuals). Then, going over each keyword file, general elimination trends (Table 1) were identified for each keyword.

Finally, a total of 20,358 usernames was collected.

Next, tweets by these usernames had to be collected for further analysis. To ensure that further algorithms and analyzes were not limited to heightened anxiety arising due to work-from-home during COVID-19, tweets from the identified users were collected over 3 periods of time—May–August 2019, May–August 2020, and

Table 1 Elimination patterns for usernames

Keyword	Raw no. of usernames	Elimination patterns	Final no. of usernames
Company	3142	• Usernames/names with "company" • Tweets using structure of pronoun + company (e.g. "keep me company")	2439
Corporate	3090	• Usernames/names with "corporate"	2059
Job	3082	• "Good job", "great job", "con job"	2112
Manager	3192	• Since "manager" is used in sports contexts, all such tweets were removed • "File manager" and "task manager" • Tweets with "fantastic manager" (news report during data collection)	1753
My Industry	3127	• Tweets containing phrases associated with the entertainment industry	2459
Product	3047	• Tweets for product reviews and promotions ("review" and "check out") • Tweets about skincare products (mostly reviews or promotions) • "Coupon" or "deal"	1728
Recruit	3134	• Since "recruit" is often used in context of college sports recruits, all related tweets were removed	1397
Salary	3032	• "Actor", "cricketer", or "NFL" • Tweets containing "TTS" (trend)	2317
Sales	3057	• Tweets with keywords about music	2032
Software	3134	• "eBay", "Amazon", "Flipkart"	2062

January-April 2021. The Tweepy API (library for communication between Python and Twitter) was used to collect tweets. A script was run for 32 days, and >1 million tweets were collected.

However, not all users had sufficient Twitter activity from all time periods, so minimum 30 tweets were needed per period. Finally, data were retained from 8598 users.

3.2 Pre-Processing

Once the Twitter data were read into Python, pre-processing was conducted. All Unicode characters, emoticons, and links were replaced with an empty string to ensure that these special characters/symbols did not contribute to model training or sentiment analysis.

3.3 Sentiment Analysis

For each user, 3 values were calculated separately for each period. These 3 values were (i) positive sentiment score, (ii) negative sentiment score, and (iii) work anxiety score.

The positive sentiment and negative sentiment scores were calculated using the TextBlob (Python library for common NLP tasks) sentiment analysis algorithm.

For the work anxiety score, some words associated with work anxiety, adapted from JAS-15, were identified, likely to be used by people suffering from work anxiety. Each time a pair of words was used in a tweet, the raw work anxiety score was incremented, and finally, the raw work anxiety score was divided by total number of words by user to obtain final work anxiety score. Whilst this was a primitive approach before a machine learning approach, it helped in understanding the linguistic potential of JAS-15.

1. "always worry" AND "work"
2. "minor matters at workplace" OR "minor matters at work" AND "worry"
3. "workplace" OR "job" OR "work" AND "restricting my capacities for achievement"
4. "suffer" OR "miserable" OR "ill health" OR "bad health" OR "not feeling well" OR "unfairness" OR "avoid" AND "workplace"
5. "work" AND "nervous" OR "my health" OR "suffer" OR "miserable" OR "ill health" OR "bad health" OR "not feeling well" OR "unfairness" OR "avoid"
6. "anxiety" OR "suffer" OR "miserable" OR "ill health" OR "bad health" OR "not feeling well" OR "unfairness" OR "avoid" AND "job"
7. "problem" AND "superiors" OR "boss" OR "colleagues"

Thereafter, for each period, the Pearson correlation coefficient was calculated between negative sentiment or positive sentiment and work anxiety.

3.4 Data Labelling

After conducting sentiment analysis, data from some users were manually read and labelled on the severity of work anxiety expressed on a scale of 0 to 1, according to JAS-15. The adaptation of JAS-15 for identifying tweets is shown below (Table 2).

Table 2 JAS-15 for Twitter

S. No.	Presence in Tweet/Account
1.	User tweets about missing work purposely or avoiding the workplace
2.	User tweets about haste to leave the workplace
3.	User tweets about specific events from the workplace that cause anxiety
4.	User tweets about problems with colleagues
5.	User tweets about problems with superiors
6.	User tweets about problems with colleagues at least 5 times over two weeks
7.	User tweets about a direct correlation between work and health
8.	User tweets about a direct correlation between workplace and health
9.	User tweets about health and work separately, but constantly
10.	User tweets specifically about work anxiety
11.	User constantly tweets about ambition and growth at the workplace
12.	User tweets about uncertainty at work
13.	User constantly tweets about work, at least five times over two weeks
14.	User tweets about the effect of work on family life
15.	User constantly tweets about work, at least seven times over two weeks

Ground truth train data were collected and labelled from 40 random users, test data were collected and labelled from 15 random users, and data from 20 users (with high work anxiety indexes) were used in creating the token bag (bag of words).

Bag of words contained 50 most common words (no stop words) from those 20 users.

3.5 Model Training

A support vector machine (SVM) algorithm was used for creating the work anxiety regressor, trained on 6 feature vectors. SVM was chosen given the small dataset (with less noise), high prediction accuracy, support for kernels, easy implementation, and memory efficiency. The dimensions of the JAS-15 were used to identify 5 features. The training and testing data also took the same feature vectors into account since they were labelled according to the JAS.

The 6 feature vectors included

Table 3 Correlation between sentiments and work anxiety

	2019	2020	2021
Positive sentiment and work anxiety score	0.045	−0.236	−0.235
Negative sentiment and work anxiety score	0.177	0.454	0.581

1. No. of times workplace mentioned (*stimulus-related anxiety*)
2. No. of times colleagues are mentioned (*social anxiety and cognition of mobbing*)
3. No. of times health is mentioned (in relation to *health- and body-related anxieties*)
4. No. of times ambition is mentioned (in relation to *cognition of insufficiency*)
5. No. of times deadline/pressure is mentioned (in relation to *job-related worrying*)
6. No. of times words from bag of words used

   A corpus was created for each feature vector and then values found.

Thereafter, an SVM regressor was trained and tested.

4 Results—Sentiment Analysis

The Pearson correlation coefficients, across the 3 time periods, between scores have been shown in Table 3.

It can be seen that generally, there is a low negative correlation between positive sentiment and work anxiety, whilst there is a moderate positive correlation between negative sentiment and work anxiety. This is consistent with the researchers' expectations as work anxiety is more strongly correlated with negative outcomes.

The correlation between negative sentiment and work anxiety became stronger between 2019 and 2021, which could be due to rising negative sentiment amidst COVID-19 and expected increase in work anxiety due to general increase in uncertainty.

5 Results—SVM Algorithm

The trained SVM regressor achieved a root mean squared error (RMSE) of 0.2 and an accuracy of 83%, which indicates that whilst the algorithm might be successful in predicting work anxiety in a certain range, it needs more data for higher accuracy and lowered error.

6 Discussion

There are several limitations to this study that might have acted as sources of error.

Firstly, the labelling of ground truth and test data was relative, due to the relative nature of the JAS-15 itself, and subject to human error. Furthermore, only 40 users were used in testing and 15 users in training, which is a small sample size. To appropriately train and evaluate, at least 100 users should have been used in testing and training alike.

There were limitations in the sentiment analysis as well. As was seen with username collection, a lot of irrelevant data had been collected. Whilst it was attempted to remove most irrelevant data, it could not be ensured that all of it was removed, and thus, lowered noise in the dataset could not be guaranteed. Furthermore, other algorithms besides TextBlob could have been used, as TextBlob only has an accuracy of ~56% [12].

Future directions could include training the algorithm with more data or implementing an app with a GUI to enable user understanding of work anxiety as well.

7 Conclusions

The hypothesis was supported in that specific linguistic markers (through the dimensions of the JAS) can be used to predict levels of work anxiety, based on JAS-15, as there is a measurable difference in language use between a healthy individual and an individual suffering from some level of work anxiety. Twitter, therefore, serves as a good source of data to predict a corporate professional's level of work anxiety.

References

1. Borg MG, Riding RJ, Falzon JM (2006) Stress in teaching: a study of occupational stress and its determinants, job satisfaction and career commitment among primary schoolteachers. Educ Psychol 59–75
2. Brockmeyer T, Zimmermann J, Kulessa D, Hautzinger M, Bents H, Friederich H-C, Herzog W, Backenstrass M (2015) Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. Front Psychol
3. Choudhury MD, Gamon M, Counts S, Horvitz E (2013) Predicting depression via social media. AAAI
4. Dormann C, Zapf D, Isic A (2002) Emotionale arbeitsanforderungen und ihre konsequenzen bei call center-arbeitsplätzen. Zeitschrift für Arbeits- und Organisationspsychologie 201–215
5. Hobson J, Beach JR (2000) An investigation of the relationship between psychological health and workload among managers. Occup Med 518–522
6. Linden M, Muschalla B, Olbrich D (2008) Die job-angst-skala (JAS). Ein Fragebogen zur Erfassung arbeitsplatzbezogener Ängste. Zeitschrift für Arbeits- und Organisationspsychologie 126–134

7. Muschalla B, Heldmann M, Fay D (2013) The significance of job-anxiety in a working population. Occup Med 415–421
8. Muschalla B, Linden M, Olbrich D (2010) The relationship between job-anxiety and trait-anxiety—a differential diagnostic investigation with the job-anxiety-scale and the state-trait-anxiety-inventory. J Anxiety Disord 366–371
9. Razavi T (2001) Self-report measures: an overview of concerns and limitations of questionnaire use in occupational stress research. University of Southampton School of Management
10. Schmalbach B, Kalkbrenner A, Bassler M, Hinz A, Petrowski K (2020) Psychometric properties of a short version of the job anxiety scale. BMC Med Res Methodol 87
11. Sczesny S, Thau S (2005) Gesundheitsbewertung versus Arbeitszufriedenheit: Der Zusammenhang von Indikatoren des subjektiven Wohlbefindens mit selbstberichteten Fehlzeiten. Zeitschrift für Arbeits- und Organisationspsychologie 17–24
12. ES S (20 July 2021) Sentiment analysis in Python: TextBlob versus vader sentiment versus flair versus building it from scratch. Retrieved from Neptune blog: https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair