Sentiment Analysis on Interview Transcripts: An application of NLP for Quantitative Analysis

Manojkumar Parmar

Robert Bosch Engineering and Business Solutions Private Limited
Bangalore, India
manojkumar.parmar@in.bosch.com

Bhanurekha Maturi

Cyient Limited
Hyderabad, India
bhanu.rekha@cyient.com

Jhuma Malik Dutt

Independent Researcher
Bangalore, India
jhuma.dutt@gmail.com

Dr. Hrushikesh Phate

Galaxy Care Hospital
Pune, India
dr.hrushikesh.phate@carehospitals.com

Abstract—One-on-one interviews and the analysis of their transcripts is the most common way researchers get into depth to obtain detailed insights. These insights are highly subjective and lack objectivity. We demonstrate in this paper a method and a use case to bring objectivity to this analysis. We present the use of NLP to generate sentiment analysis and perform various quantitative techniques. This analysis is useful in deriving insights by finding patterns and building a simple linear model to explain the variation in sentiment pattern. We also present a view about the usage of this technique for the effective and optimal time usage by researchers to learn maximum from outlier interviews.

Index Terms—NLP, QDAP, sentiment analysis

I. INTRODUCTION

In social sciences, a primary research methodology such as one-on-one interviews constitute a widely used technique to derive meaningful insights and draw comprehensive conclusions [1]. In the business scenario dominated by the influx of technology, customer discovery is accomplished through one-on-one interviews to derive at the requisite product features, pricing, and launch strategies [2]. Personal interviews are a treasure trove of information. Once transcribed, these interviews help in providing qualitative analyses. However, such analyses are subjective and draw heavily from the unconscious biases of the authors or researchers. That apart, the learning from every new interview diminishes at a high rate and is not an efficient use of the valuable time of the researchers.

We have used a known technique of sentiment analysis from the domain of natural language processing (NLP) to provide objectivity to the interpretation of personal interviews and derive significant insights [3]. We have developed this technique at the intersection of social research, NLP, and analytics, which has no known precedence. We have applied sentiment analysis on the answers provided by a person to a set of questionnaire, averaged the value to derive the final sentiment of the interview. We have curated a dataset of interviews and performed an exploratory analysis on it to improve the understanding. With this analysis, we comprehend the pattern of sentiment about the questions and the demographics. This relationship provides further insights which are

very difficult to assimilate by simple qualitative analysis. Our method can be used to reduce the burden on senior researchers by flagging non-conforming interviews to the learnt pattern. These flagged interviews can be studied in detail to provide enhanced understanding at minimal effort.

II. BACKGROUND AND PROBLEM STATEMENT

That the millennials are unique is a general understanding. We conducted a research to provide a detailed description of their uniqueness in the Indian scenario by accessing views of managers and millennials from multiple industries. First, we performed literature survey to understand the current viewpoints and later designed separate questionnaires to assess the views of millennials, managers and HR managers. We conducted 18 interviews in total across five industries. Their views are captured using semi-structured interviews. The transcripts are processed to generate qualitative views. The qualitative analysis provides with a detailed understanding of millennials. There is a lack of a standardised way to ascertain the qualitative analysis [4]. We discussed and realised a need for quantitative analysis to add the needed objectivity and standardisation. We present here our research method and associated quantitative analysis as a solution to the stated problem statement using NLP techniques.

III. RESEARCH METHOD

Based on the literature survey, we created two sets of questions one for millennials and the other for managers to figure out their respective views. In total, ten questions were administered to each participant. The answers were audio recorded, transcribed against each question along with other variables - industry of operation, gender, experience, a category of input (millennial, senior manager, HR manager). These steps were carried out manually. On recorded transcripts, we performed a sentiment analysis each of the answers using NLP. If the answer to any question is not available, then it is considered as missing data, and missing data is not imputed. The average sentiment value was derived for each participant which forms the baseline data for further analysis. We created

a baseline dataset by performing sentiment analysis on the answers provided by each participant, and combined them with other demographic variables.

We investigated this dataset using descriptive methods, and visualisation to improve our understanding. This understanding further led us to perform hypothesis-building and validation to unearth the factors influencing the sentiments and establish a statistical relationship. With the proven statistical relationship, we built a linear model to explain variance in average sentiment along with correlation coefficients. At each stage, we obtained new observations and insights.

IV. DATASET GENERATION

In total, the authors conducted 18 interviews across five industries covering nine females and nine males. The dataset consists of eight senior managers, two HR managers and eight millennials. Table I represents the distribution of interviews over the industries, gender and position held in the organisation. The questions and their answers are available in the dataset [5]. We processed the transcribed answers to produced stemmed version which provides essential base words from the answers. Each base word is then classified as positive or negative using a pre-built library. For research, we have used the QDAP library of positive and negative words [6]. Based on the classified positive and negative words, the difference in count is calculated and normalised. This normalised variable is used as the sentiment of the answer. The sentiment analysis provides a single number between -1 and 1. The normalised variable can take value from -1 to 1, where -1 indicates complete negative sentiment and +1 indicates a complete positive sentiment. A variable with value 0 is considered as neutral sentiment. We have answers to 10 questions, and hence averaging is applied to derive average sentiment of the interview from the available answers.

V. DESCRIPTIVE ANALYSIS

In this section, we use descriptive statistics to deepen our understanding of the dataset. Table I represents the entire distribution of interviews over demographic variables. We use this table to provide a summary of the dataset. Based on sentiment analysis, table II represents a weighted average of sentiment analysis over demographic variables, which we use to build visualisation. The descriptive summary and visualisation are helpful in identifying influences and building hypothesis.

From table I, it is amply clear that the dataset has equal gender representation with nine females and nine males. Millennials and senior managers are well represented. However, the representation of HR managers is poor. From an industry standpoint, automotive/IT, construction, and geospatial/IT are well represented. It is important to note that the millennial group is heavily represented by females while senior manager group is heavily represented by males.

The results of sentiment analysis are available in Table II. Table II represents weighted average sentiment based on the number of participants. It is clear that the overall sentiment is positive. However, the degree of positiveness differs for

specific industry, gender, and position. The millennial group is the most optimistic among all. The female group, in general, is more positive than the male group. Automotive/IT industry demonstrates the highest positiveness among others, post discarding medical and banking group due to poor representation. Figure 1 represents visualisation of table II.

The visualisation provides all data points with differentiated markings for industry, gender and millennials while plotting the average sentiment over industry experience. From figure 1, we can infer that with increasing industry experience, the average sentiment is reducing but staying positive.

VI. HYPOTHESES RESULTS

From our descriptive analysis and visualisation, we infer a relationship which needs to be validated by a statistical means using hypothesis testing. Hypothesis validation is critical to converting observations to concrete knowledge with statistical significance.

The sentiment variation has a dependency on the gender, industry, industry experience, and millennial status. These multiple dependencies are analysed using ANOVA technique [7]. Table III provides results of ANOVA and it is clear that industry and millennial status have significant statistical relationship with sentiment considering 95% confidence interval. We do not have sufficient data samples to do detailed hypothesis validation with respect to gender in the given industry and vice-versa, and other similar comparisons.

From figure 1, it is apparent that as experience increases, sentiment reduces. We observed that the correlation is -0.745746 with the p-value of 0.0004 using Pearson's product-moment correlation method [8]. With this result, it is clear that correlation is statistically significant but the ANOVA results in table III are stating otherwise. This result implies that experience is not a dominant variable which explains the variation of sentiment despite strong correlation.

VII. LINEAR MODEL

To extend the understanding, we built the linear model to demonstrate the variance explanation and correlation coefficient. The industry is a categorical variable, and millennial is a binary variable. These two variables together can take six different positions and can explain 72.15% variation in the average sentiment [9]. Table IV provides the value of coefficients with their significant t-values and p-values.

A. Interpretation of Model

The intercept 0.1737 represents the average sentiment value of the automotive/IT industry non-millennial candidate. The millennial status increases average sentiment by 0.1078. For the construction industry, non-millennial and millennial status reduces by 0.005. Similarly, geospatial/IT industry, non-millennial and millennial status reduces by 0.081.

TABLE I
NUMBER OF INTERVIEWS

Industry	Millennial		Sr. Manager		HR Manager		Total		Total
	F	M	F	M	F	M	F	M	
Automotive/IT	1	1	0	1	1	0	2	2	4
Banking	0	1	0	0	0	0	0	1	1
Construction	2	0	1	2	0	0	3	2	5
Geospatial/IT	3	0	0	3	0	1	3	4	7
Medical	0	0	0	0	1	0	1	0	1
Total	6	2	1	6	1	2	9	9	18
	8		7		3		18		

TABLE II
WEIGHTED AVERAGE SENTIMENT

Industry	Millennial		Sr. Manager		HR Manager		Total		Total
	F	M	F	M	F	M	F	M	
Automotive/IT	0.316	0.323	0	0.123	0.149	0	0.233	0.223	0.228
Banking	0	0.4	0	0	0	0	0	0.4	0.4
Construction	0.29	0	0.216	0.132	0	0	0.265	0.132	0.212
Geospatial/IT	0.166	0	0	0.122	0	0.108	0.166	0.118	0.139
Medical	0	0	0	0	0.264	0	0.264	0	0.264
Total	0.233	0.361	0.216	0.125	0.207	0.108	0.225	0.176	0.2
	0.265		0.138		0.174		0.2		

Fig. 1. Weighted average sentiment analysis plot

TABLE III
HYPOTHESES RESULTS - DEMOGRAPHIC VARIABLES

Variable	Degree of Freedom	Sum of Sq	Mean of Sq	F Value	p-Value	Remarks
Industry	4	0.0740	0.0185	9.439	0.0020	Significant
Millennials	1	0.0455	0.0455	23.217	0.0007	Significant
Gender	1	0.0002	0.0002	0.076	0.7888	Non-significant
Experience	1	0.0007	0.0007	0.339	0.5734	Non-significant
Residual	11	0.0196	0.0020			

TABLE IV
LINEAR MODEL COEFFICIENTS

Variables	Coefficients	t statistics	p-value	Significance	Possible value
Intercept	0.1737	7.519	7.05e-06	Yes	
Millennials	0.1078	5.172	0.00023	Yes	0-1
Industry- Construction	-0.0050	-0.180	0.86013	No	0-1
Industry - Geospatial/IT	-0.0810	-3.127	0.00874	Yes	0-1

TABLE V
LINEAR MODEL RESULTS AGAINST ACTUALS

Industry	Millennial Status	Indicated Average sentiment	From data (average sentiment)	Deviation
Automotive/IT	Yes	0.2815	0.319	0.0375
	No	0.1737	0.136	-0.0377
Construction	Yes	0.2765	0.29	0.0135
	No	0.1687	0.16	-0.0087
GeoSpatial/IT	Yes	0.2006	0.166	-0.0346
	No	0.0927	0.118	0.0253
Average		0.199	0.1982	-0.0008

B. Model Results and Comparison

Table V represents the model results and its comparison with actual data to ascertain and understand the working of the model. The linear model provides an accuracy of 72.15% on the collected data. The model predicted sentiment for representative types for given industry and millennial status is a very close approximation of data available. The deviation between model predicted and actual varies in the range of -0.0377 to 0.0253 only.

VIII. DISCUSSION

This quantitative method to analyse the interview transcripts is a newly developed method. In total, we conducted 18 interviews across five industries covering nine females and nine males. The dataset consists of eight senior managers, two HR managers and eight millennials. The sentiment analysis performed on all interviews provides positive values; however, Millennials are highly optimistic compared to managers, which provide evidence about the optimistic nature of millennials. The ANOVA results provide a clear indication about the dependence of variation of sentiment based on industry and being millennial. Additionally, as experience increases, the sentiment is reducing, but it is not conclusive enough as it is not able to stand the statistical test as a whole. This result is evidence of general understanding that millennials and managers see millennials through different lenses. The

built linear model which provides a detailed understanding of sentiment variation can explain 72% variation in sentiment just based on industry and millennial status variable.

The quantitative analysis slightly differs from the qualitative analysis and points out one interesting aspect that within an industry, millennials do view themselves differently from managers and across the industry there is a variation in the degree of conformity to this relationship. The millennials have a different impact across different industries, and they need to be understood in the context of the industry to take advantage of the optimism brought by millennials in the organisation for transformation. At the same time, the organisation should also look at millennials more positively and try to accommodate them for future growth.

IX. FUTURE WORK

This method has potential usage in validating genuineness of the interviews and also to identify outliers for detailed investigation. This method has potential to reduce the burden on researchers as they can learn more about the subject from outlier interview transcripts which are not confirming general trend derived from the linear model. The model presented here can be further improved by using weighted average sentiment analysis where the weight of each questions sentiment is provided. The weight of each question can be derived using Analytic Hierarchical Process. This method is extended to

perform the question-wise detailed analysis for better insights. The authors have the plan to increase the size of the data points and also perform the weighted average analysis.

X. CONCLUSION

We developed an innovative technique to apply NLP to derive the average sentiment of interview transcripts. We demonstrated the use of quantitative analysis to derive meaningful insights from the sentiment analysis leading to hypothesis validation and building a linear model. We built a linear model with two simple variables to explain 72% variation in the sentiment. The method we have developed can be used to find out outlier interviews to improve learning for researchers in a highly efficient way.

REFERENCES

- [1] P. Burnard, "A method of analysing interview transcripts in qualitative research," *Nurse Education Today*, vol. 11, no. 6, p. 461466, 1991.
- [2] S. Blank and B. Dorf, *The startup owner's manual: The step-by-step guide for building a great company*. BookBaby, 2012.
- [3] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [4] H. R. Bernard, "Qualitative data, quantitative analysis," *CAM Journal*, vol. 8, no. 1, pp. 9–11, 1996.
- [5] M. S. Parmar, M. Bhanureka, J. M. Dutt, and H. Phate. (2018, January) Interview dataset. [Online]. Available: https://www.dropbox.com/s/pimuj63c0afdbz3/Interview_Database.xls
- [6] B. Goodrich, D. Kurkiewicz, and T. Rinker, "Package 'qadp'," Tech. Rep., 2018. [Online]. Available: <https://cran.r-project.org/web/packages/qdap/qdap.pdf>
- [7] J. M. Chambers, T. J. Hastie *et al.*, *Statistical models in S*. Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove, CA, 1992, vol. 251.
- [8] L. Statistics, "Pearson product-moment correlation," *Lund Research, Inc*, 2013.
- [9] B. H. Margolin and R. J. Light, "An analysis of variance for categorical data, ii: small sample comparisons with chi square and other competitors," *Journal of the American Statistical Association*, vol. 69, no. 347, pp. 755–764, 1974.