Assessment of Communication Styles in a Corporate Setting

Aquila Khanam **Keerthana Kumar** **Leighanne Hsu** **Rishina Tah**

Columbia University

Abstract

Language and communication styles in corporate settings are greatly affected by social context. Drawing from previous work on corporate email analysis, we aimed to use the Enron email corpus to detect formality and the presence of requests in emails, to analyze how they relate to gender and positions of power.

Two binary classifiers were trained using the dataset of 400 emails with a range of different kernels, parameters and features. The best request classifier (76.8% accuracy) and best formality classifier (78.0% accuracy) were applied to a set of gender and position annotated emails. It was found that women are slightly more formal than men, while both are more formal than informal. Both men and women are more formal in communication via email than informal. All positions tend to use formality while communication through emails in a work place. It is also seen women make more requests, while men make more non-requests. Superiors write more emails than subordinates, and most positions send fewer requests than non-requests.

There were some limitations in the experiment, including the large number of unannotated gender and position information. In the future, a politeness classifier can be created to further analyze requests and formality.

1 Introduction

Social context influences language, mannerisms, and other facets of communication. Linguists and sociologists have analyzed the influence of social context and societal expectations on behavioral modifications. One such behavioral modification is politeness, which is regarded as an important aspect of expression. It is interesting to observe the social networks within large organizations to analyze how different groups interact with each other. Interaction between peers would be less formal than interaction between a superior and a subordinate. Gender may also play an important role in determining the degree of formality and requests. This interaction can be best analyzed by looking at email interactions between individuals of a particular company. A variety of natural language processing and machine learning techniques can be applied to decipher emails and extract the degree of formality and number of requests sent. In this project, we aim to detect formality and presence of requests in emails, and analyze how they relate to gender and positions of power. This helps in observing patterns of behaviour and understanding differences in language and communication styles in a corporate settings.

2 Past Work

An integral part of the detection of politeness in emails is the assessment the formality in the tones of the sender and recipient. The formality could be affected by several factors, including social distance, relative power and the weight of the imposition (Peterson et al., 2011). The weight of an imposition is computed by analyzing requests.

Brown and Levinson (1987) formulated the idea of requests as negative politeness, but prior to that, Lakoff (1973) and Brown and Levinson (1978) explored various minimization strategies i.e. methods that speakers employ to reduce the intensity imposition of the request, such as indirect speech and apologies.

There has been plenty of prior work regarding differences in language usage between men and women. Researchers have found that women are more likely to foster personal relations (Deaux et

al., 1987; Eagly et al., 1984) and share concerns and support others (Boneva et al., 2001), while men tend to talk about activities and communicate in the interest of social position (Tannen, 1991).
Extrapolating from previous work by Mohammad and Yang (2011), we think gender and position of the sender and recipient may play a factor in formality and politeness of requests of an email. Thus, we intend to incorporate gender differences into our classifier to see if the genders of the sender and recipient changes the formality or tone.

3 Research Questions

By creating two separate classifiers and applying it to a set of corporate emails, we aimed to answer the following questions:
Gender:

- Who sends more requests?
- Who is more formal?
- Who sends more formal requests?

Positions of power:

- Do superiors send more requests?
- Are subordinates more formal?
- Who sends more formal requests?

4 Dataset

The Enron corpus was used as the dataset for this project. This dataset contains emails sent by employees of the corporation, to other employees as well as external clients. This is the most popular and largely used publicly available datasets for analyzing corporate emails and departmental networks. A modified version of the dataset created by Prabhakaran et al. (2014) was used, as this corpus contained gender annotations for all employees, along with position information. A portion of the preprocessing was already done by separating useful fields of an emails (e.g. to, from) and storing it in collections in MongoDB.
Using this dataset, 12,061 emails were manually extracted. Further preprocessing was done by removing blank emails with no content, and separating the body from other parts of the email thread such as previous replies and forwards. There were 10,343 valid emails for request analysis and 11388 for formality analysis to work with after completing the preprocessing steps.

5 Experimental Setup

Peterson et al. (2011) separated a set of 400 emails from the Enron corpus, and annotated each as informal or formal, and 0 and 1 for contains request and does not contain request, respectively. Using this annotated dataset, a stratified split (based on both formality and request, separately) was performed to allocate 300 emails for training, 50 for development, and 50 for testing.

For each classifier, a baseline was created. New features were added, and a range of different classification algorithms were implemented. The classifier with the best accuracy was run on the final set of 10,343 emails for requests and 11,388 emails for formality to perform statistical analysis.

6 Classifiers

6.1 Formality Classifier

The baseline formality classifier was adapted from the Grinbox-server, which is a Google Chrome extension to perform sentiment analysis on emails on Gmail. This extension had a formality classifier built into it to separate out the business emails from personal emails. The training, development and test sets were built from the 400 emails annotated by Peterson et al. (2011). The baseline classifier gave an accuracy of 50% on the test set.

Classifier	Type	Style
1	Baseline	50%
2	Customized training set	54%
3	Trigrams	40%
4	Relevant features	40%
5	Modified annotations	66%
6	Iterative model	72%
7	Balanced dataset	78%

Table 1: Different formality classifiers

The baseline classifier used a bag-of-words and naive bayes classifier to compute formality. It produced poor results because the words in the model were not tailored to the Enron corpus. This improved the classifier accuracy to 54% (Classifier 2).
Since this is not very high either, we tried using trigrams (40% accuracy), added more features like message length to Classifier 2 (60% accuracy) using the empirical data that formal emails are longer.
Finally, we decided to try a different approach and

made two important changes:
Using an iterative model for training, which is to analyze the emails in the development set that were tagged incorrectly and then retraining the classifier on the augmented training+development set.
Creating a balanced data set. One important observation was that since there were almost double the number of formal training files than informal ones, the classifier was biased towards tagging all emails as formal. The final training set had 119 formal and 119 informal emails and the accuracy achieved on the test set was 78%.
These are the final features used for the formality classifier:

Feature	Weight Proportion
formal_punctuation =False	inform:formal = 4.3 : 1.0
abbreviations =True	inform:formal = 3.0 : 1.0
informal_punctuation = True	inform:formal = 2.7 : 1.0
msg_length = True	inform:formal = 2.1 : 1.0
polite = True	inform:formal = 2.0 : 1.0
misspelled = False	inform:formal = 1.8 : 1.0
slurs = True	inform:formal = 1.7 : 1.0
informal_punctuation = False	inform:formal = 1.5 : 1.0
slurs = False	inform:formal = 1.4 : 1.0

Table 2: Different formality classifiers

6.2 Request Classifier

For request classification, we used the dataset from Peterson et al. (2011). We started by extracting trigram feature vectors from the training, replacing numbers and dates with a unique token. We then trained various classifiers on the training set and tested them on the development set to obtain an accuracy. We used the same majority baseline as Peterson et al. (2011), which classifies all emails as non-request emails. This provided an accuracy of approximately 61%, as expected from the stratified split.
We started by using trigrams as features, as this was the best classifier found by Peterson et al. To obtain the trigrams, we used the tokenizer and n-gram functions in the Natural Language Toolkit (NLTK). We first tried trigram classification in two different ways: using frequency feature vectors and using binary feature vectors. For frequency, we counted the number of times each trigram appeared in training and development. The best accuracy we obtained across several different classifiers was 48%, which was much worse than our baseline. For the binary method, we simply checked whether or not each trigram existed in each set to create a binary feature vector. As this performed much better, we used the binary feature vectors for further testing of various classifiers.
We started with support vector machines for classification. Linear, polynomial, and radial-based function (RBF) kernels were tried out. We varied C in powers of 10 from 10^0 to 10^8. We varied γ from 10^{-2} to 10^{-4}, where $\gamma = 1/2\sigma^2$. For the polynomial kernel, we tried degrees of 2 and 3. On the binary feature vectors, the best parameter values were $C = 100000$ and $\gamma = 0.001$ using an RBF kernel, which gave a 64.71% accuracy, marginally better than the baseline.
Additionally, we tried other classifiers, including K-Nearest Neighbors and Random Forests, but these had approximately 64-66% accuracy as well, so we decided to stick with SVMs.
In an attempt to improve our accuracy, we also tried two methods of feature selection, following the example set by Peterson et al. (2011). We started with selecting trigram features that appeared at least 5 times in the training corpus, and tried again selecting trigram features that appeared at least 10 times in the training corpus. These raised our accuracies slightly, with ¿5 resulting in the better accuracy, as expected, but not to any significantly higher value. We then tried using the chi-squared selection method, which checks for inter-factor dependence. We used this again with our SVM classifier and selected only the k best features. We tried varying values of k between 100 and 10000 and obtained our highest accuracy of 76.47% using $k = 1500$.
Apart from experimenting with different classifiers, additional features were added to the existing trigram feature set, based on the work of Lampert et al. (2010). These included message length (number of characters), message length (number of words), number of non-alphanumeric characters, presence of question words (could, would, who, what, where, when, why, how, which, can, may), and the presence of a question mark. For these we tried only the SVM classifier using an RBF kernel. We varied the parameters C be-

tween 1000 and 10000000 and γ between 0.01 and 0.0001, but this did not affect the accuracy much, if at all. Our base accuracy with the added features was 65%. We also performed feature selection using the chi-squared method described earlier. We varied the number of selected features between 100 and 1500, and our best results were obtained at 350 features. However, this was still found to decrease the accuracy to 72.55%, lower than our trigram-only feature model. It is possible that in addition to the trigram features, these extra features simply added noise to the classification, or they were too similar, as the trigrams captured punctuation and individual words within them as well. As a result, the additional features were removed from the final classifier.

As the SVM classifier using the top 1500 chi-squared selected binary trigram features performed best, we chose this as our final classifier. We ran this classifier on our test set and obtained an accuracy of 65%. It is not surprising that this number is lower than the accuracies in the table above. This may be due to overfitting on the development set we were working with.

A summary of some of the classifiers and parameters is below.

Classifier	Type	Parameters	Feature Selection	Accuracy
1	Baseline	--	--	61\%
2	SVM (RBF) with frequency trigram features	C = 100000; gamma = 0.001	n >= 5	48\%
3	SVM (RBF) with frequency trigram features	C = 10000000; gamma = 0.01	n >= 10	62\%
4	SVM (RBF) with binary trigram features	C =100000 ; gamma = 0.001	chi2 values, top 1500 features	76\%
5	SVM (RBF) with binary trigram and added features	C =1000, 10000, 100000, 10000000; gamma = 0.001	--	65\%
6	SVM (RBF) with binary trigram and added features	C = 100000; gamma = 0.001	chi2 values, top 350 features	72\%
7	K-Nearest Neighbors	k = 13	n >= 5	65\%
8	K-Nearest Neighbors	k = 20	n >= 5	66\%
9	Random Forests	Trees = 10	n >= 5	66\%
10	Random Forests	Trees = 20, 50, 1000	n >= 5	65\%

Figure 1: Different Request Classifiers

7 Results and Analysis

7.1 Gender Analysis

The results for formal and informal emails are given in the figures below.

We have a total of 6160 emails which are formal and 5228 emails which are informal. This is understandable as the data set consisted of both enron and work related formal emails and informal conversations between employees. As with request emails, it can be seen that women send more emails than men in a corporate setting. Women tend to be more formal in writing emails than men. It is interesting to note that when we look at absolute numbers women send out higher number of informal emails as well. This is because total number of emails that women send out is twice as much as men and hence there is a large number of emails classified as informal and written by females.

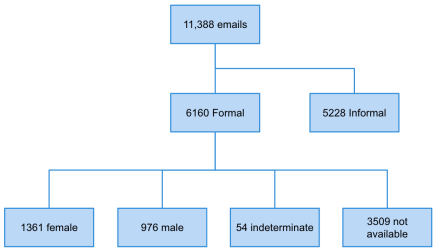

Figure 2: Gender Analysis for Formal emails

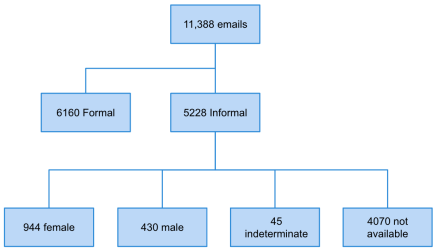

Figure 3: Gender Analysis for Informal emails

The counts obtained for males and females were normalized by the number of males and females present. The results for emails containing requests and non-requests are given in the figures below.

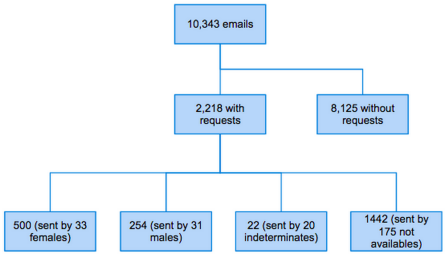

Figure 4: Gender Analysis for Requests

Finally, the requests and non-requests were classified as formal or informal to assess communication styles based on gender. The resulting fig-

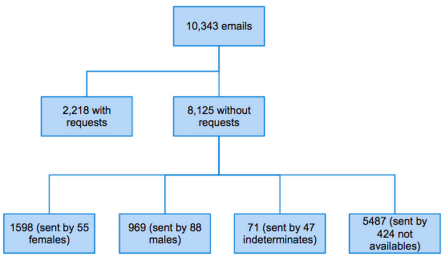

Figure 5: Gender Analysis for Non-Requests

ures are given below. We see that most requests are classified as formal. This seems right as people tend to be more formal while requesting something in a corporate setting. Once again, we see that more women tend to write formal requests than men.

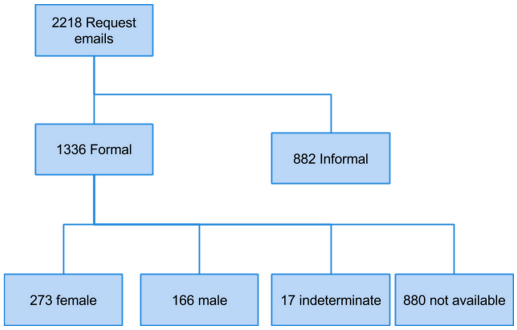

Figure 6: Formal Requests

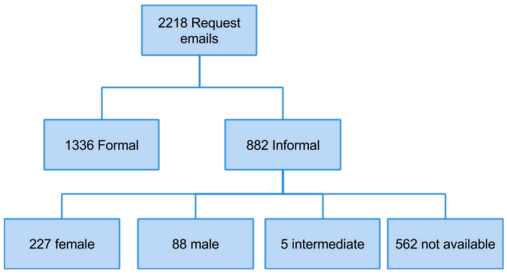

Figure 7: Informal Requests

In non-requests, most non-request emails are classified as informal. This can be attributed to the fact that most emails not consisting of requests might be informal conversation between employees.

Since the training set was limited and the test set is missing gender annotations on a large portion of the data, these results should not be taken as necessarily conclusive, but rather as a suggestion of a potential direction of further study. These are the

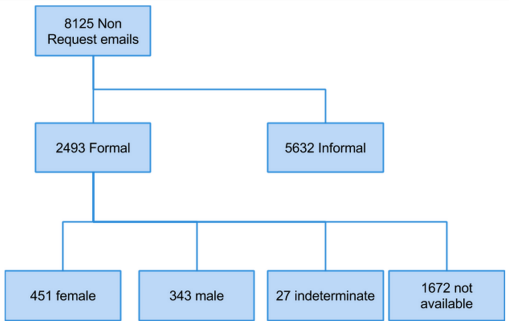

Figure 8: Formal Non-Requests

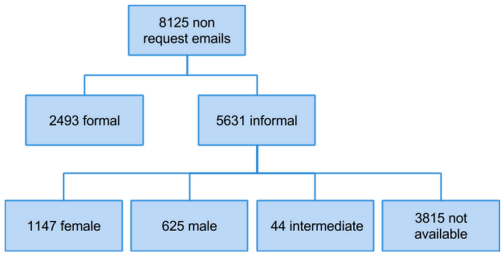

Figure 9: Informal Non-Requests

conclusions that can be drawn from the preliminary analysis:

- Women send more total emails.
- Men and women are more formal than informal at work.
- In proportion of the emails they write, men are more formal than women.
- Women send more non-requests. On average, they send 15 requests and 29 non-requests per person.
- Males send 8 requests and 11 non-requests on average
- Men send more formal non-requests than women

The greater number of non-request emails is not surprising since non-requests comprise 78.56% of the dataset. However, it is interesting that women seem to send nearly twice as many non-requests as requests. One possible reason is that many of these are social and not business-oriented emails. A larger or better-annotated data set could help confirm these findings.

7.2 Position Analysis

Like with the gender data, we were also missing lots of position tags. This results in some positions only having one or two people tagged. Thus, many of the positions may only have one person listed, which is unlikely for most positions. Therefore, just as before, it is necessary to only use our results as a direction of guidance, and future study with a larger or better data labeled set is needed.
We ran analysis of the formal and informal emails with respect to position. There was a lot of missing data and according to our results all of the annotated positions sent out formal emails. For instance, below are some positions of power that tend to be **formal** in their emails in a corporate setting.

- VP
- Trader
- Lead
- Business
- Managing

The counts obtained for different hierarchies and positions of power were normalized by the number of tagged employees per position. The results for emails containing requests is given in the figure below. It is seen that VP and Asst Counsel send the most requests, by a large margin. Managing directors, Executive VPs and lawyers follow, but with a much lesser number. The number of requests sent by the remained of the employees is smaller and closer in number (smaller variance). Non-senior positions such as specialists, contractors and commercial directors are higher up in the graph, with much fewer requests sent.

Similar analysis was done for emails without requests, as shown below. It can be seen that the spread of counts is much wider than in the case of requests. VP and Asst General Counsels once again send the most number of non-requests, with Lawyers and Executive VPs following more closely behind. Non-senior positions like traders and analysts send more non-requests than senior positions such as Sr Vice President and Sr Director.

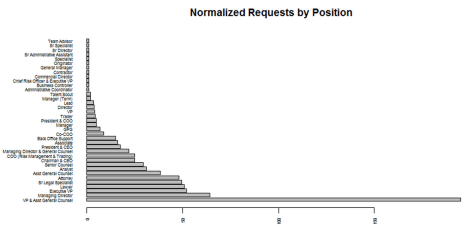

Figure 10: Normalized Requests by Position

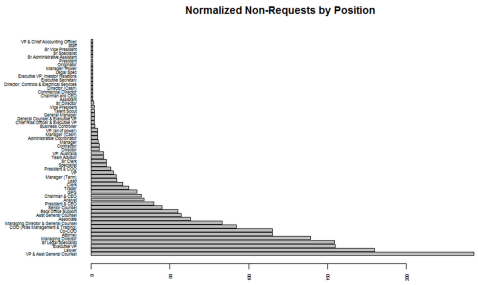

Figure 11: Normalized Non-requests by Position

A detailed table of figures for each position can be found in the appendix.

8 Conclusion

Based on our analysis, the following conclusions can be drawn.
Gender:

- Women are slightly more formal than men
- Both men and women are more formal than informal
- Women write approximately twice as many emails as men
- Women make more requests than men
- Men make more non-requests than women
- Women tend to make slightly more formal requests than men

Position:

- Superiors write more emails than subordinates
- All positions apart from Senior VP send fewer requests
- All positions are more formal than informal in a corporate setting

Other observations:

- Superiors write more emails than subordinates
- More number of requests are formal in nature than informal
- Most non-request emails are informal in nature

9 Limitations and Future Work

The goal of this project was to analyze formality and the number of requests in emails - as they both relate to politeness in a corporate setting - and analyze the differences across genders and positions of power. We were able to meet the objectives and answer the research questions, given the allotted time frame. However, more extensive analysis can be done with the data obtained. There were some limitations that we faced, which did not allow us to provide a deeper analysis. For example,

- Approximately 70% of the emails from the Enron corpus had missing gender information. Of the remaining 30%, 5% were marked as unknown in the database.
- Approximately 20% of the emails from the Enron corpus had missing position information. Of the remaining 80%, 20% were marked as unknown in the database.
- The training set of 300 emails was not large enough to provide a thorough classifier, as ngrams from only these 300 emails were used as features for machine learning. There may have been a large number of ngrams in the analysis set that were not seen by the classifier.
- The hierarchy of positions and departments within the organization are not known. We were able to visualize the hierarchy only in cases where senior is clearly mentioned in the position title.

In order to get a better classification accuracy in the future, we propose the following:

- Incorporate more extensive features such as the presence of full names in the signature, layout of the email, etc.
- Experiment with a wider range of classifiers/kernels with different parameters
- Use a larger annotated training set for building the classifiers

In addition, a politeness extension can be incorporated to solidify the analysis of the relationship between politeness and formality/requests.

10 Combined Work

This paper was written by all four group members. The coding and analysis portions of the project were distributed as follows:

- Formality Classifier: Aquila and Rishina
- Request Classifier: Keerthana and Leighanne

11 References

A. Agarwal, A. Omuya, A. Harnly, O. Rambow, *A Comprehensive Gold Standard for the Enron Organizational Hierarchy*. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012.

A. H. Eagly and V. J. Steffen, *Gender stereotypes stem from the distribution of women and men into social roles*, Journal of Personality and Social Psychology, 1984.

A. Lampert, R. Dale, C. Paris, *Detecting Emails Containing Requests for Action*, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pp. 984992, 2010.

B. Boneva, R. Kraut, and D. Frohlich, *Using e-mail for personal relationships*, American Behavioral Scientist, 2001.

C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, C. Potts, *A Computational Approach to Politeness with Application to Social Factors*, ACL 2014.

D. Tannen, *You just dont understand: women and men in conversation*, Random House, 1991.

K. Deaux and B. Major, *Putting gender into context: An interactive model of gender-related behavior*, Psychological Review, 1987.

K. Peterson, M. Hohensee, F. Xia, *Email Formality in the Workplace: A Case Study on the Enron Corpus, Proceedings of the Workshop on Languages in Social Media*, pp. 86-95, 2011. J. Otterbacher, *Inferring gender of movie reviewers: exploiting writing style, content and metadata*, Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM, pp. 369378, 2010.

M. Thelwall, D. Wilkinson, and S. Uppal, *Data mining emotion in social network communication: Gender differences in myspace*. J. Am. Soc. Inf. Sci. Technol., 61:190199, 2010.
P. Bramsen, M. Escobar-Molano, A. Patel, R. Alonso, *Extracting Social Power Relationships from Natural Language*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.
P. Brown and S. C. Levinson. 1978. *Universals in language use: Politeness phenomena*, Questions and Politeness: Strategies in Social Interaction, pp. 56311, 1978.
P. Brown and S. C. Levinson, *Politeness: Some Universals in Language Usage, Cambridge University Press, 1987.* R. Lakoff. *The logic of politeness: Minding Your Ps and Qs*. Proceedings of the 9th Meeting of the Chicago Linguistic Society, pp. 292305, 1973.
S. M. Mohammad and T. Yang, *Tracking Sentiment in Mail: How Genders Differ on Emotional Axes*, Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp. 7079, 2011. V. Prabhakaran, E. Reid, O. Rambow, Gender and Power: How Gender and Gender Environment Affect Manifestations of Power, EMNLP, 2014.

Appendix

Position information for emails with request

TotalRequests	TotalInPosition	NormalizedRequests	
VP & Asst General Counsel	195	1	195
Managing Director	257	4	64.25
Executive VP	52	1	52
Lawyer	51	1	51
Sr Legal Specialist	198	4	49.5
Attorney	96	2	48
Asst General Counsel	153	4	38.25
Analyst	31	1	31
Senior Counsel	207	7	29.57142857
Chairman & CEO	25	1	25
COO (Risk Management & Trading)	25	1	25
Managing Director & General Counsel	22	1	22
President & CEO	35	2	17.5
Associate	16	1	16
Back Office Support	15	1	15
Co-COO	9	1	9
GPG	7	1	7
Manager	20	4	5
President & COO	5	1	5
Trader	22	5	4.4
VP	38	9	4.222222222
Director	34	9	3.777777778
Lead	18	5	3.6
Manager (Term)	2	1	2
Talent Scout	2	1	2
Administrative Coordinator	1	1	1
Business Controller	1	1	1
Chief Risk Officer & Executive VP	1	1	1
Commercial Director	1	1	1
Contractor	1	1	1
General Manager	1	1	1
Originator	1	1	1
Specialist	1	1	1
Sr Administrative Assistant	1	1	1
Sr Director	1	1	1
Sr Specialist	1	1	1
Team Advisor	1	1	1

Figure 12: Position Information for Requests

Position information for emails without request

TotalNonrequests	TotalInPosition	NormalizedNonrequests	
VP & Asst General Counsel	486	2	243
Lawyer	180	1	180
Executive VP	155	1	155
Sr Legal Specialist	772	5	154.4
Managing Director	835	6	139.1666667
Attorney	230	2	115
Co-COO	115	1	115
COO (Risk Management & Trading)	92	1	92
Managing Director & General Counsel	166	2	83
Associate	63	1	63
Asst General Counsel	572	10	57.2
Back Office Support	55	1	55
Senior Counsel	719	16	44.9375
President & CEO	40	1	40
Analyst	134	4	33.5
Chairman & CEO	64	2	32
GPG	29	1	29
Trader	214	9	23.77777778
Clerk	20	1	20
Lead	130	8	16.25
Manager (Term)	16	1	16
VP	241	17	14.17647059
President & COO	25	2	12.5
Specialist	19	2	9.5
Sr Clerk	19	2	9.5
Team Advisor	8	1	8
VP: Australia	8	1	8
Director	136	26	5.230769231
Contractor	5	1	5
Manager	43	10	4.3
Administrative Coordinator	4	1	4
Manager (Cash)	4	1	4
VP (all of power)	4	1	4
Business Controller	10	4	2.5
Chief Risk Officer & Executive VP	2	1	2
General Counsel & Executive VP	2	1	2
General Manager	2	1	2
Talent Scout	2	1	2
Vice President	2	1	2
Sr Director	9	5	1.8
Assistant	1	1	1
Chairman and CEO	1	1	1
Commercial Director	1	1	1
Director (Cash)	1	1	1
Director: Controls & Electrical Services	1	1	1
Executive Secretary	1	1	1
Executive VP: Investor Relations	1	1	1
Legal Spec	1	1	1
Manager: Power	1	1	1
Originator	3	3	1
President	1	1	1
Sr Administrative Assistant	2	2	1
Sr Specialist	1	1	1
Sr Vice President	1	1	1

Figure 13: Position Information for Non-Requests