National College of Ireland

The Prepared Executive: A Linguistic Exploration of Executive Answers in the Q&A Section of Earnings Calls

MSc Research Project
Data Analytics

Tom Donoghue
x16103491

School of Computing
National College of Ireland

Supervisor: Dr. Ralf Bierig

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Tom Donoghue
Student ID:	x16103491
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Ralf Bierig
Submission Due Date:	16/08/2017
Project Title:	The Prepared Executive: A Linguistic Exploration of Executive Answers in the Q&A Section of Earnings Calls
Word Count:	6634

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	16th August 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

The Prepared Executive: A Linguistic Exploration of Executive Answers in the Q&A Section of Earnings Calls

Tom Donoghue
x16103491
MSc Research Project in Data Analytics

16th August 2017

Abstract

Earnings calls provide a stage for senior executives to present information to financial analysts. The interactive setting of the call's question-and-answer section offers an opportunity to explore executives' style of answering analyst questions. This study examines linguistic features in executive answers and investigates how these features might be proxies for detecting executive preparedness. The motivation is to discover hitherto unknown facets of the executive through their language. A sample of call transcripts is used to obtain the executive answers. A domain expert participated in labelling a sample to provide a ground truth. Word lists and document similarity models, using natural language processing and latent similarity analysis techniques, measure the proxy features of uncertainty, avoidance and repetition. The feature models' results are combined to yield the executive's answer as prepared or unprepared. Model performance is evaluated through accuracy, precision, recall, F_1 measure and Cohen's Kappa. The results show the models' performance is inconsistent detecting these feature in comparison to the domain expert. However, the models do find small amounts of the sough-after features, which may warrant a deeper inspection of the executive's answer to discover what is actually being said.

1 Introduction

Corporate earnings conference calls enable organisations to present pertinent operational and financial information to the investment community. The calls involve senior executives from the presenting organisation and the financial analysts representing the investor community. The executives usually comprise the chief executive and chief financial officers and may include regional or other senior managers. The organisation may add their investor relations (IR) manager to facilitate the calls progress. Most of the call is concerned with the dispersal of prepared financial information, which is communicated by the executives. The latter part of the call is consumed with the question and answer session (Q&A) in which the analysts participate in an interactive discourse with the executives. It is the Q&A that is of interest and in particular the executive answers (EA). Executives undergo rehearsals to practise answering the likely questions which they may

face on conference calls (Larcker and Zakolyukina; 2012). The preparation and impact of IR scripted Q&A responses as Lee (2016) finds that executives adhere to the script to avoid divulging negative information. Part of the training is to prepare the executive to provide short and concise answers. They are also expected to have the capability to field questions regarding the company's performance and their role in executing the company's strategic business plan. The purpose is to place the executive in an optimal position of preparedness to answer the analysts' questions. The motivation for this study is the opportunity to discover executive preparedness through the exploration of distinct linguistic features. Our focus is on the executive, as according to the domain expert, less is known about them in relation to information on investment analysts. What type of linguistic patterns appear in the EA? How might these patterns show the executive as uncertain and caught unawares by a question and hence indirectly unprepared? How might the findings expose facets in "executive speak" which signals the need for a deeper inspection of a company's methods of communication. Hence, the research question and hypotheses:

Research Question: How might the combination of uncertainty, avoidance and repetition indicate the degree of preparedness in executive answers to analysts' questions?

Hypothesis 1: EA that comprise more vague language are perceived as more uncertain.

Hypothesis 2: EA that comprise more distancing language are perceived as applying avoidance tactics.

Hypothesis 3: EA that comprise similar language are perceived as being more repetitive.

Hypothesis 4: EA that show a combined degree of uncertainty, avoidance and repetition are perceived as being more unprepared.

The linguistic features selected are: uncertainty ((Burgoon et al.; 2016) and (Szarvas et al.; 2012)); avoidance (McDonald; 1990) and repetition ((Pennebaker and King; 1999) and (Palmieri et al.; 2015)). The context diagram in Figure 1 illustrates the process.

The drinks industry is chosen for this study due to its brand diversity, merger and acquisition activity, broad market reach and vociferous executives. 36 earnings call transcripts (2457 EA) from 8 companies are obtained from call transcripts captured from seekingalpha.com.[1] Using this raw dataset, EA extracts from the Q&A section are produced for downstream exploratory and observational linguistic analysis. To provide external validity to the analysis, a domain expert (with comprehensive IR knowledge in the drinks sector) provides a ground truth by participating in an experiment to label an EA preparedness sample (n=208, drawn from the 2457 EA). This annotated sample provides the core classified dataset for this study. The domain expert agreed to participate in the study having reviewed the author's project proposal. The expert recognised that the exploration of executive language is a challenging domain and was keen to be involved. A contribution to the area of research is made by building on the linguistic analysis of earnings call transcripts using adapted and possibly untried combination of word lists (Loughran and McDonald (2011); Pennebaker and King (1999) and Tausczik and Pennebaker (2010)) and Latent Semantic Analysis (LSA) (Duran et al.; 2010). Securing the participation of a senior domain expert with experience in the drinks industry and the

[1] http://seekingalpha.com/search/transcripts [Accessed 16 August 2017]

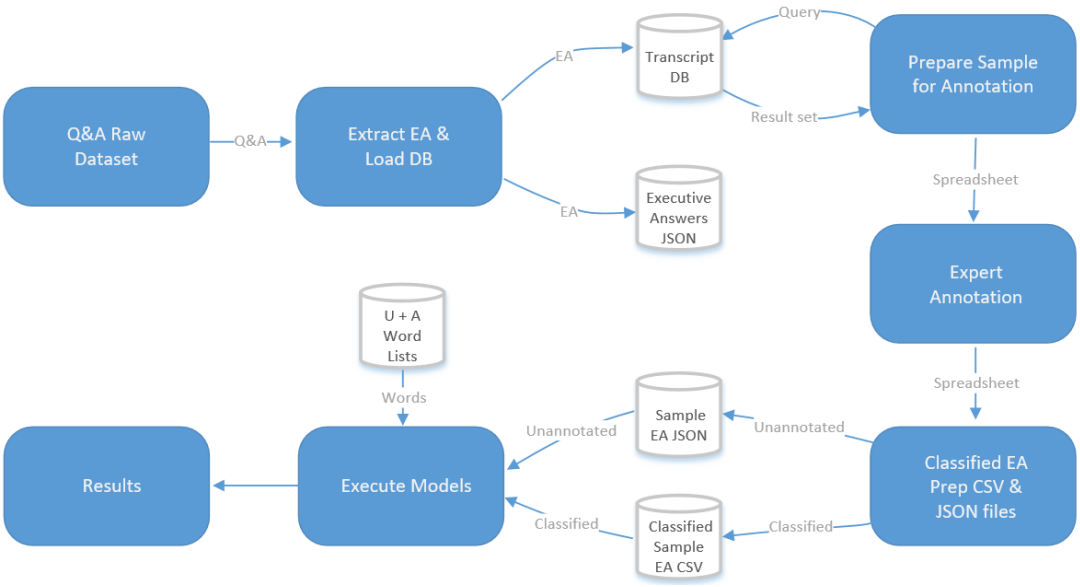

Figure 1: Process Context Diagram

production of the annotated EA dataset is a valuable contribution, and as far as the author is aware this type of data is not generally available. Finally, the addition of the pre-processed transcripts may offers scope for future research to be conducted.

The following section covers the related work in this domain. Section 3 covers the methodology and implementation, addressing the data, model selection, EA sample experiment and model execution. Section 4 covers the results and limitations. Section 5 concludes and describes areas for future research.

2 Related Work

The analysis of earnings calls attracts the interest of researchers from accounting and finance and psychology backgrounds. The research tends to examine the impact on the market following the earnings call discourse. A review of the literature is covered across the following dimensions: linguistic analysis of the Q&A section, linguistic treatment of the selected features in this research and the use of textual analysis word lists and machine learning techniques.

2.1 Q&A Section of Earnings Calls

Research in this domain includes: deception, tone, structure, stock price and executive spontaneity. Accessing the Q&A discourse data is achieved through access to call transcripts on third-party financially focused websites. The use of audio is scarce, but Mayew and Venkatachalam (2012) obtain audio files streamed via a third party subscription. Earnings call audio is also used in Burgoon et al. (2016) who use a small sample of executive audio clips (n=1,114). The authors agree that a larger sample size drawn from a wider range of companies would be more representative. The authors also note that their executives are of different gender. Time and resources constrain data collection in their analysis. The same constraints are experienced in this study, with its small sample

dominated by North American and European English speaking males. Many conference calls are streamed live via webcast and may subsequently be available via the company's website (Hobson et al.; 2012). Video could provide a new dimension of audio-visual cues for exploration of participant behaviour and exhibited styles (Price et al.; 2012). The video analysis of earnings calls as Burgoon et al. (2016) and Hobson et al. (2012) concur has yet to be widely undertaken. Focus in this study is concerned with the linguistic style of EA in the Q&A section; which is more free flowing and hence more likely to give rise to the unexpected (Larcker and Zakolyukina; 2012). Similar studies in which Matsumoto et al. (2011) and Burgoon et al. (2016) concur that the Q&A section caters for a greater amount of information dispersal based on the interactivity with analysts in attendance. The behaviour of executives to act on impulse or to appear phlegmatic is as Lee (2016), Hollander et al. (2010), Matsumoto et al. (2011) and Larcker and Zakolyukina (2012) find the Q&A section accounts for higher degrees of impromptu divulgence. The power-play between executives and analysts in the Q&A section as Palmieri et al. (2015) find the executives protect and explain their position on questioning. The authors also find that analysts afford executives space rather than exercising an aggressive mode of questioning. The above appears to support exploration of EA linguistic features in the Q&A.

2.2 Linguistic Features of Preparedness

In the discussion of the Larcker and Zakolyukina (2012) study of deception by Bloomfield (2012) suggests that misleading executives are capable of avoiding disingenuous linguistic cues. The author and, as already mentioned, Larcker and Zakolyukina (2012) concur that executives are trained and rehearse answers to possible questions. The analysts invited to participate on the call is under the control of the organisation's IR team which is substantiated by Mayew (2008) and Brockman et al. (2015). Executives are either protecting themselves from inquisitive analysts asking difficult questions or executives have the ability to sidestep tricky questions. How, given executive training, is it possible to detect their level of preparedness through potential proxy features? Careful review of the literature has presented the following promising features as potential candidates for models in this domain.

Uncertainty through the executive's use of language construed as unclear, equivocal or vague (Burgoon et al.; 2016) and (Szarvas et al.; 2012). How might the executive be caught off-guard by the question and hence their style of language lacks a relaxed manner and succinctness? Regarding deception Burgoon et al. (2016) hypothesise that the language of prepared and unprepared utterances is different, and use hedging and uncertainty as a feature. People tend to communicate in a language that is more cautious when they are uncertain regarding a subject as Tausczik and Pennebaker (2010) mention the heightened use of "tentative" and filler words. Uncertainty in responses may be more pronounced during periods of adverse company performance as Matsumoto et al. (2011) suggest. However, if executives have been sufficiently prepared for questions in advance one should expect little discernible variance in the levels of uncertainty and hence the executives degree of preparedness. Hesitation may also indicate uncertainty as Thomas et al. (2015) index uncertainty using hesitations and unplanned utterances. The authors find that the unpreparedness consists of the inadvertent use of hesitation and pauses and hence may indicate executive uncertainty.

Avoidance is a second feature where the executive appears to evade the question by: answering a question with a question; referring to someone else to answer the question;

deferring a succinct response as it is dependent on some future event and deviation by changing the subject or answering a different question (McDonald; 1990). The verb tense used in the answer may show that the answer is being shifted into the future, this enables the executive to state what positive future actions will be conducted. For example, it is easier to describe what you will do rather than being constrained by what you do or have done. In review of the Gunsch et al. (2000) study of political ads Tausczik and Pennebaker (2010) mention similar findings associated with verb tense and personal pronoun usage. The same authors also note the use of tentative language and filler words in avoiding the question (e.g. "sometimes", "sort" and "seems"). The use of first person and deflection away from one's self as Larcker and Zakolyukina (2012) find (in reference to Vrij (2008) and Newman et al. (2003) findings), are similar to Bloomfield (2012) and concur with the thoughts and expectation of this study's domain expert. Over-elaborate responses and answers which meander could also indicate avoidance of the question. The words and phrase selected by an executive's answer represent their thoughts and actions, but also tell the audience about their level of confidence (Thomas et al.; 2015).

Repetition is a third feature to explore. The executives use of causal words which according to Pennebaker and King (1999) may indicate their consistent need to explain a given situation by rephrasing the subject (e.g. "reason", "because" and "why"). The use of discrepancy type words (e.g. "would", "should" and "could") used in reasserting the point or subject. The authors mention insightful words associated with comprehension (e.g. "know" and "realise"). The repetition of words in the answer may indicate that the executive is feeling unsure of getting their point across. Executives may repeat points when they perceive they are unpersuasive. In using executive argumentation to explain information value of earnings calls, Palmieri et al. (2015) find they offer an opportunity for executives to persuade analysts to accept their standpoint. Should the executive appear to be unable to sustain a persuasive argument this may erode their level of confidence leading to an increase in repetition. Increased information may be beneficial to the analyst community and investors. However, the reiteration or inflation of previous statements may dilute the value of information or create an unintentional loss of impact (Bozanic and Thevenot; 2015). Repetition may be recognised through pronounced reuse of subjects, phrases, themes or sentence similarity ((Duran et al.; 2010) and (Lee; 2016)).

2.3 Language, Text Analysis, Words & Machine Learning

The language of the Q&A section and attempting to determine the linguistic style and how to categorise these styles as Pennebaker and King (1999) suggest is an arduous task. One method which the authors suggest is the use of a word- or dictionary-based approach. The authors derive a linguistic style by following Allport (1961) study of behaviour, and suggest that the words used during the performance of communication indicates a specific style. This links to the behaviour of impulse, reactionary or involuntary remarks made on the conference calls (Lee (2016); Larcker and Zakolyukina (2012)). Recognition of the above does little to counter the loss of context in the use of word based analysis, which motivates Pennebaker and King (1999) to implement their own textual analysis computer system called "LIWC" (Linguistic Inquiry and Word Count) which comprises multiple lexicons and psychological taxonomies to provide improved meaning and understanding of text. Textual analysis can lure one into trouble as Loughran and McDonald (2016) indicate loss of context when extracting information and failure to classify a label correctly may confound exploration and hypotheses. In a survey of corporate disclosure Li (2010)

suggests that risks include: a naïve approach to the textual analysis process,[2] perceiving causality due to association, failure to include or check an attribute which contributes to the impact on a dependent variable. Although Loughran and McDonald (2016) examine written financial documentation, the approach to text analysis in this study is similar. The authors also cover the use of phrases, word lists and dictionaries, warning that lists of words and dictionaries may become adrift of their context once extracted from their original setting. Both lexical methods, however may preclude presumptions made by the researcher and scale well using software. Reuse of these standard dictionaries as Loughran and McDonald (2016) suggest enable researchers to begin with a common set of classifiers with which to follow and recognise their use in previous studies. Table 1 indicates the standard lists and dictionaries used in the earnings call space. Supplementing and tuning world lists created by Loughran and McDonald (2016) as Allee and Deangelis (2015) and Brockman et al. (2015) confer, assists classification when aligned with a financial vocabulary.

Name	Description	Used in	Comment
Henry (2008)	Positive/Negative word list created for earnings press releases	Price et al. (2012) Doran et al. (2012) Davis et al. (2015)	Used for assessing tone & sentiment. Has a relatively small word count
Harvard General Inquirer	Derived from Harvard IV-4 dictionaries	Price et al. (2012) Loughran and McDonald (2011) Tetlock (2007)	Used widely in financial text analysis particularly to assess tone of newspaper copy
Diction	Positive/Negative word lists	Davis et al. (2015)	Large taxonomy of positive & negative words. Favoured by researchers in accounting
Loughran and McDonald (2011)	Positive/Negative word list better suited to a financial lexicon based on 10-K filings	Allee and Deangelis (2015) Davis et al. (2015) Brockman et al. (2015) Chen et al. (2014) Mayew and Venkatachalam (2012)	Wide range business words in 6 word lists and a master dictionary

Table 1: Word lists and dictionaries used in textual analysis, based on Loughran and McDonald (2016) and Kearney and Liu (2014)

Naïve Bayes and Support Vector Machines are common machine learning methods used in text classification ((Guo et al.; 2017); (Sun et al.; 2017) and (Loughran and McDonald; 2016)). These supervised learning methods require labelled data to train the model. The model is then applied to a test set of data to measure its performance in predicting the class label. The creation of labelled data is a challenge which Ittoo et al. (2016) recognise, and point to the requirement for an acceptable or agreed foundation dataset(s) with which to prime the model. Using Latent Dirichlet Allocation (LDA) (Blei et al.; 2003) which is an unsupervised learning model to detect topics in the document space may preclude the need for discourse to be labelled (Sun et al.; 2017). Latent Semantic Analysis (LSA) is used to measure sentence similarity as Duran et al. (2010) investigate words with similar meaning appearing in different sentences. LSA uses singular

[2]this is an evolving domain which requires the science of economics to provide a backbone to hypotheses.

value decomposition to create a lower rank approximation of the original term document matrix. For a value k representing a new k-dimensional space, k should be in the low hundreds according Manning (2015). Hence the term document matrix C is remapped to a lower dimensional space C_k, which results a m x n matrix in which each document (row) has k (columns). Using cosine similarity between two documents provides a score between 0 and 1. The closer to 1 the more similar the documents. Cosine similarity is a common method used in natural language processing (NLP) to measure document similarity (Loughran and McDonald; 2016) and (Lee; 2016). Although LSA is a precursor to later models (LDA and probabilistic LSA) it may preclude issues of synonyms (similar meaning different word) and polysemy (differing meaning same word) in word based models. However, LSA may suffer from low performance when using small documents (Loughran and McDonald; 2016). The need for pre-processing of the raw text data to ensure its readiness as input for textual analysis remains. The pre-processing comprise the following activities: breaking the text into separate entities (i.e. tokenisation), deletion of unwanted stop words (e.g. 'the', 'an', 'a', 'it') which have little bearing on downstream text analysis, changing text to lowercase, reducing words back to their base (i.e. stemming), calculating the term frequency inverse document frequency (*tf–idf*) how often a term appears in a document and inverse document frequency.[3] This process remaps the document to a vector space which is more appropriate for use in information retrieval (e.g. exploring document classification, clustering and similarity) ((Manning; 2015); (Uysal and Gunal; 2014); (Ghiassi et al.; 2013) and (Blei et al.; 2003)).

Research by Ittoo et al. (2016) and Kumar and Ravi (2016) indicate that there is a reawakening in the field of textual analytics which seeks to extend NLP and machine learning into deep learning using neural networks. The authors advocate future research to explore evolving techniques which ease textual analysis to produce meaningful results which are comparable. The classification of documents using deep learning are beyond the scope of this study, but the use of software (e.g. gensim[4] and LSA (Řehůřek and Sojka; 2010)) which enables document similarity comparisons to be made in a lower dimensional space, could offer a simpler alternative. This study builds on the current research through linguistic analysis of earnings call transcripts using a possibly untried combination of word lists (Loughran and McDonald; 2011); (Pennebaker and King; 1999) and (Tausczik and Pennebaker; 2010) and LSA (Duran et al.; 2010) to explore the preparedness of EA with, as far as known, a new feature mix and annotated EA sample dataset.

3 Methodology and Implementation

The objective is to explore EA from the call transcripts using textual analysis techniques, extract the required features which when combined may indicate the level of unpreparedness. The process is illustrated in Figure 2. The executive and analyst utterances from the raw Q&A file (which are provided from an earlier work by the author) are read, identified, sequenced saved to a JSON file. Each sequence file is read, the EA are extracted to create a new EA JSON file which contains all answers for a given executive on a specific call. The diagram in Figure 3 shows the relationships of the call components. The concept of each entity is clearly described in Table 2. The 36 call transcripts are collec-

[3]log of the number of documents of interest divided by the count of documents encapsulating the term

[4]http://radimrehurek.com/gensim/ [Accessed 16 August 2017]

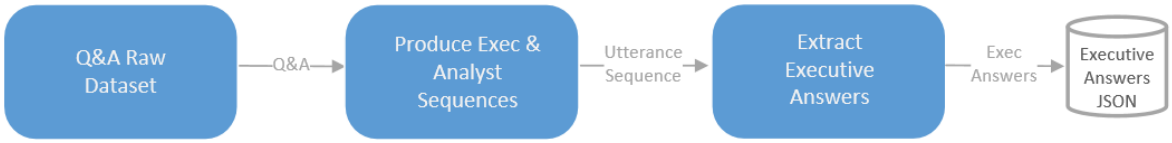

Figure 2: Obtaining Q&A Executive Answers

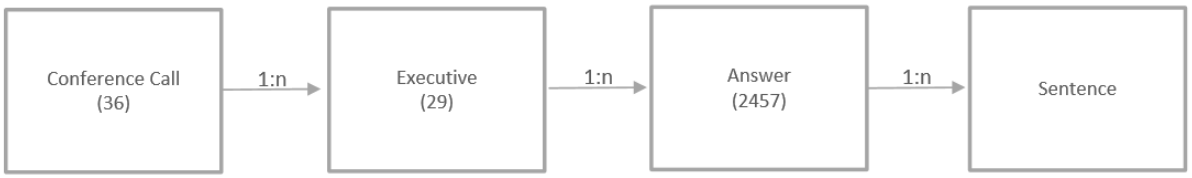

Figure 3: Conference Call Entities (counts in brackets)

Entity	Description
Conference call	Call transcripts which only use the Q&A are extracted. The scope is to explore the EA hence analyst questions are excluded.
Executive	Usually the CEO and CFO. However, additional senior executives and an IR executive may be included.
Answer	An utterance by a given executive in response to an analyst's question. Here, each answer represents a document.
Sentence	The number of grammatical sentences that comprise an answer.
Token	The lowest level of granularity, the terms which make a sentence.

Table 2: Conference Call Entity Descriptions

ted from 8 companies in the drinks industry. The calls are based on the quarterly calls available for this sector. The companies are a mixture of European and North American producers of beers, wines and spirit brands. The language on the calls and subsequent transcripts is English (US spellings). Each conference call / executive relation represents a corpus of EA (i.e. the documents) and each EA represents a document. For example, as shown in Table 3 Company "A" has 2 executives, "Exec 1" who utters 2 answers and "Exec 2" uttering 1 answer. It is the executive's answers (i.e. the documents) that form the basis of the exploration in this study.

Company	Executive	Answer
Company A	Exec 1	"Yes, it would appear that the market reacted in a way that..." "And on your second question, I would expect..." ...
	Exec 2	"I agree with that response. A larger share..." ...

Table 3: Conference Call EA Examples

3.1 Model Selection

Testing the hypotheses requires models appropriately grounded in the literature. The basic models are selected from disparate studies but are adapted and combined in an unprecedented way to measure uncertainty, avoidance, repetition and unprepared for this study. Uncertainty uses a word list model, based on Loughran and McDonald (2011) and similarly followed in Lee (2016) and Brockman et al. (2017) gauging tone. The model measures uncertain tokens as a proportion of total answer tokens. Each token t in each answer j is used to look up the list of uncertain words. If a match is found the uncertain token count u is incremented. The uncertainty score for that document is calculated as a value between 0 and 1, subtracting the value from 1 provides the uncertainty score. Hence providing a quantitative measure of uncertainty for hypothesis 1.

$$Uncertainty_j = 1 - \frac{\sum_{i=1}^{n} t_{i,j} - \sum_{i=1}^{n} u_{i,j}}{\sum_{i=1}^{n} t_{i,j}} \tag{1}$$

The Avoidance model also uses a word list model, but this study extracts the LIWC word lists widely used in linguistic analysis in this domain (e.g. Lee (2016); Larcker and Zakolyukina (2012) and Li (2008)) and measures the avoidance tokens as a proportion of of total answer tokens.

$$Avoidance_j = 1 - \frac{\sum_{i=1}^{n} t_{i,j} - \sum_{i=1}^{n} a_{i,j}}{\sum_{i=1}^{n} t_{i,j}} \tag{2}$$

where a is the count of avoidance tokens. The model is augmented by the addition of the future tense as a percentage of present and past tense which Li (2008) measures

$$FuturePastPresent_j = log\left(\frac{\sum_{i=1}^{n} 1 + f_{i,j}}{\sum_{i=1}^{n} 1 + p_{i,j}}\right) \tag{3}$$

where f is the count of future tense verbs, p is the count of past and present verbs. The author also uses self versus you and other, where s is self and y is you and other.

$$SelfYouOther_j = log\left(\frac{\sum_{i=1}^{n} 1 + s_{i,j}}{\sum_{i=1}^{n} 1 + y_{i,j}}\right) \tag{4}$$

The use of first person and deflection away from one's self as Larcker and Zakolyukina (2012) and Bloomfield (2012) concur with the thoughts of the domain expert. Total avoidance is obtained by this study's combination of the above results and provides a quantitative measure of avoidance for hypothesis 2.

$$TotalAvoidance_j = (FuturePastPresent_j - SelfYouOther_j) + Avoidance_j \tag{5}$$

The repetition model measures how similar each EA sentence is to each other sentence in that answer. Similarity measurement is used in Lee (2016) to compare executive utterances between the Q&A and management discussion sections of the call. The model selected in this study uses LSA and the calculation of sentence similarity score between 1 and 0. The repetition score provides a quantitative measure of repetition for hypothesis 3.

$$Repetition_j = \begin{cases} 1 & \text{if } score_j > 0 \\ -1 & \text{otherwise} \end{cases} \tag{6}$$

The feature scores are used to assign a value of 1 or -1 indicating the presence or absence of the sought-after feature.

$$FeaturePresent_j = \begin{cases} 1 & \text{if } score_j > 0 \\ -1 & \text{otherwise} \end{cases} \tag{7}$$

The combiner model accounts for each feature's contribution to the measure of unpreparedness by summing the feature present values to indicate unpreparedness, and provides the measure for hypothesis 4.

$$Unprepared_j = \begin{cases} 1 & \text{if } \sum_{j=1}^{n} FeaturePresent_j > 0 \\ -1 & \text{otherwise} \end{cases} \tag{8}$$

With a set of models defined, the objective is investigate how they perform when run against classified EA.

3.2 Executive Preparedness Sample Labelling Experiment

Running the models against a labelled dataset provides a quantitative approach based on a ground truth. Model performance may be measured using this ground truth. An experiment is designed to label a sample of EA by the domain expert. The domain expert possesses industry knowledge and IR expertise to classify the dataset objectively to deliver the ground truth desired. Obtaining access and the participation of a domain expert adds validity to the study. The involvement of an individual with superior knowledge, proximity and familiarity with the study's target subjects is rare. Hence, their contribution of the annotated dataset elicits a unique offering in this domain. Figure 4 illustrates the creation of the sampling process.

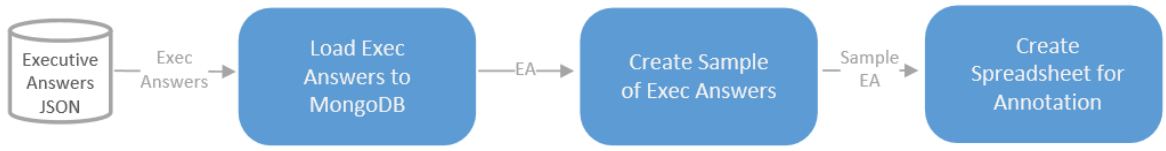

Figure 4: Creating EA Sample for Annotation

3.2.1 Data and Design

The data for the experiment is a subset of the original conference call data as described in Section 3. The subset is a random sample (n=208) of EA taken from the 36 earnings conference call transcripts. The sample's distribution is compared to population from which it is drawn to check that it is representative. There are 27 distinct executives represented in the sample and the proportion of answers appears to be representative of the population. The sample is saved to a template spreadsheet with the following columns:

- Call id – identifies the company and quarter to which the conference call relates
- Executive Answer – the answer transcript
- Executive Name – the name of the executive answering
- 4 columns are added representing the features to annotate: Uncertainty; Avoidance; Repetition and Unprepared

An answer id is added to uniquely identify each answer in the spreadsheet. The sheet is saved following removal of Call id and Executive Name columns (to avoid leading the annotator). The domain expert is given instructions via email which includes the annotation spreadsheet. Each feature is described and the method for scoring for feature presence i.e. -1 = not present; 0 = neutral; +1 = present.
Uncertainty "The language used by the executive is construed as unclear, vague, or equivocal. The answer is ambiguous or fuzzy".
Avoidance "The language indicates that the executive appears to evade answering the question. Answers that are over elaborate, abrupt or meandering."
Repetition "The executive reuses words, phrases, subjects or themes or a high perceived sentence similarity."
Unprepared "Overall score for the how unprepared the executive appears to be in answering the question."

A score is required for each feature and the spreadsheet is scored sequentially top to bottom. The raw answer text is provided verbatim, as this study is exploratory and hence the removal or obfuscation of names and exclusion of outliers may introduce bias rather than reduce it (Strang and Sun; 2016). Obtaining a gold-standard dataset together with the benefit and cost of using domain expert annotation as Ittoo et al. (2016) concur is a far from a simple activity. Usually more than one of annotator is employed to mitigate bias as Schmidt et al. (2016) find that a common consensus amongst annotators provides a better result. The study is constrained by timescale and resources and the availability of only one domain expert to annotate the dataset. However, noting this limitation, the domain experts experience is aligned to the setting and may well outweigh the use of several annotators with less experience or from other domains (but this remains an unknown).

3.2.2 Experiment Results

The domain expert scored each feature as instructed. It was noted that there could be an overlap between features and a score was to be attributed to all four features for each answer. Of the 208 answers in the sample 197 were annotated (11 answers are excluded as they are continuity utterances and 1 is a question). Figure 5 illustrates the percentages of EA across the 4 features. According to the domain expert executives are trained and prepared ahead of conference calls. However, the level of uncertainty, avoidance, repetition and unpreparedness appear to be higher than expected. There are 7 examples in the results producing intriguing combinations. 4 answers classified as uncertain, avoiding and repetitive, 3 of these 4 are also classified as unprepared = neutral, and the remaining 1 example is classified as prepared. 3 other examples are classified as unprepared when the other features indicate that they either neutral or certain, non-avoiding and non-repetitive. The domain expert suggests that a prepared executive could exhibit uncertainty and may well use avoidance tactics because they are prepared for a question. Taking an unprepared executive, it may indicate good management ability when they provide an answer on which they have drawn on knowledge about their business. Hence, they exhibit certainty, non-avoidance and non-repetition. One of the 4 hypotheses is based on the 3 features as proxies for indicating unpreparedness. The annotation results show examples where the hypothesised correlation between the proxies and the unprepared feature is reversed or neutralised. Could this mean that unpreparedness is an additional proxy and that the 4 should be combined to indicate some higher concept

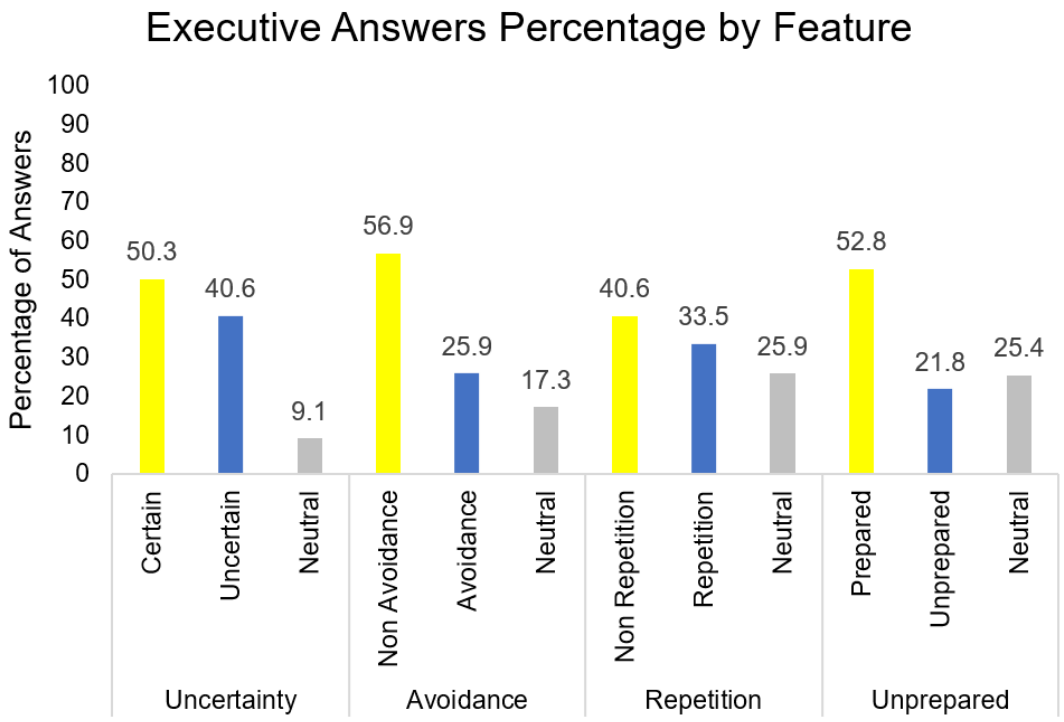

Figure 5: Annotated Results - Executive Answers Percentage by Feature

e.g. EA dexterity? This could be considered an area for future research. These cases could be outliers but it hints that the models need to be sensitive to inconsistency and opposing linguistic styles.

3.3 Model Execution

The models are run to evaluate their performance with the classified EA and the corresponding unannotaded EA. The process is illustrated in Figure 6. Using the master sample

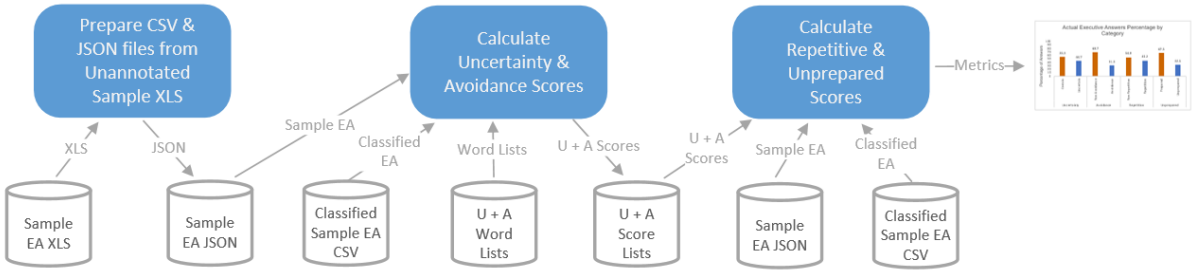

Figure 6: Model Execution Workflow

spreadsheet of EA an unannotated JSON file is created which contains the 197 unlabelled EA. A CSV copy of the 197 annotated EA is manually saved. The files together with the uncertain and avoiding word lists are input to calculate uncertainty and avoidance scores. Using the Python NLTK package the EA undergo text pre-processing (e.g. stop word removal, lowercase conversion and document tokenisation). The uncertainty and

avoidance scores are calculated for each tokenised EA using the equations (1 - 5 and 7). The EA examples annotated as feature neutral are removed as the models create dichotomous results (i.e. the feature is either present or absent, there is no in between). The results are saved out for inclusion in the repetitive and unprepared calculation. As the repetitive model uses a different architecture to the word list it is run separately. The same 197 EA input files are used. The EA are tokenised at sentence level (as sentence are being compared) and tagged with their EA id and stop words removed. Using gemsim's Python package to convert the tokenised sentences to a sparse vector space from which a lower dimensional space is created using gensim LsiModel with $k = 250$ dimensions. Which produces a matrix with 689 documents and 250 features. As mentioned above Manning et al. (2008) suggests setting k in the low hundreds yields better precision. The space is searched using the source sentence, returning the top 5 highest similarity scores (i.e. most similar to the source and excluding the source sentence). The returned sentence is checked that it belongs to the same EA, if so the sentence is marked as repetitive otherwise non-repetitive (equation 6). The previous results are loaded and the combined with the repetition results. The results for each feature are combined (equation8) and compared to the annotated sample.

3.4 Model Performance Metrics

Model classifier performance metrics are produced, using the Python sklearn.metrics package, for each model using a confusion matrix and scores for accuracy, precision, recall, F_1 and Cohen's Kappa. A confusion matrix shows the composition correct and incorrect predictions made by the model when compared to the actual classified features in the annotated EA sample. Accuracy is the percentage of correctly predicted features.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

Precision is the percentage of correctly predicted positive features of total predicted as positive or how well the model predicts the positive class.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

Recall is the percentage of correctly predicted positive features of actual positives or how much of what is actually positive is predicted as positive.

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

F_1 is a harmonic mean of the precision and recall with a score between 0 and 1. It provides a more complete view of the model (i.e. rewards when precision and recall are similar) than using precision or recall separately.

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{12}$$

Cohen's Kappa measures the agreement between the model's predictions and the actual true values whilst catering for agreement that may occur by chance

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{13}$$

where *Pr(a)* is proportion of actual agreement and *Pr(e)* is the proportion of expected agreement (Lantz; 2015).

4 Results

The results of this exploration of EA preparedness and the support for the proposed hypotheses are described below. The comparative feature percentages are shown in Figure 7, the model performance metrics appear in Table 4.

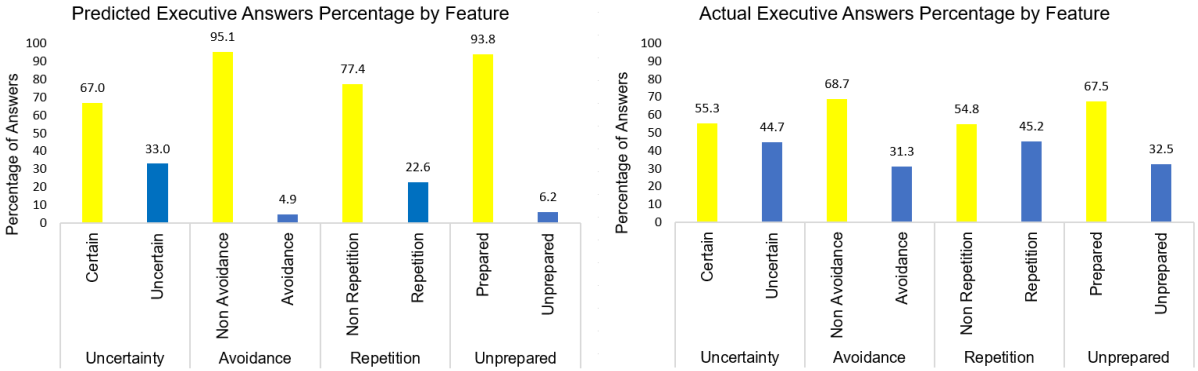

Figure 7: Predicted and Actual EA Feature Percentages

Model	Accuracy	Precision	Recall	F_1	Kappa
Uncertainty	51%	44%	33%	0.37	-0.01
Avoidance	70%	62%	10%	0.17	0.09
Repetition	55%	52%	26%	0.34	0.06
Unprepared	74%	29%	8%	0.12	0.03

Table 4: Model Performance Metrics

Uncertainty accuracy is only marginally better than a random guess at 51% with low precision 44% and recall 33% also reflected in the F_1 measure 0.37. Kappa -0.01 tends to suggest a strong disagreement between the model's predictions and the true values in the annotated sample. Kappa being slightly negative carries no more weight than score of 0. This offers weak support for the uncertainty hypothesis. Even though the model predicts 33% of EA as uncertain versus the annotated 44.7%, the domain expert's classification approach may use more than the appearance of uncertain words (from the word lists), this is an ability that the model is unable to compete with.

Avoidance accuracy is much better at 70%, precision 62% and low recall 10% and a F_1 0.17 due to the precision recall difference. Kappa 0.09 is positive but still suggests poor agreement. There is a class imbalance which favours non-avoidance. The high precision indicates that the model performs well when it recognises avoidance but the low recall is a concern. This is an example of why high accuracy and precision at the cost of low recall may indicate a model which could struggle to generalise well. The model only predicts 4.9% of EA as avoiding versus 31.3% by the expert which indicates that the expert finds more avoiding behaviour than the model is capable of detecting. To support the hypothesis a higher recall and Kappa is desirable. Detection of avoidance could be

like detecting sarcasm; the model battles to recognise the subtleties in play. Detecting avoidance requires a model which identifies the avoidance cues (veiled or otherwise) while accommodating differences based on personality and culture.

Repetition accuracy is little better than random at 55%, precision 52% a lower recall 26% and F_1 0.34. Kappa 0.06 suggests poor agreement. The model has a better balanced of accuracy and precision but recall is only correctly predicting 26% of the actual avoiding class. The model predicts 22.6% of EA as repetitive versus 45.2% by the expert, this could be due to the expert picking up on additional repetitive cues which the model is unable to perceive. EA are short documents and this may offer the model less opportunity to detect repetition. The model may also perform better than the simple word list models due to recognising synonyms and polysemy. The model offers marginal better support for the repetitive hypothesis, the precision / recall trade off might be acceptable in certain settings but the low Kappa would urge caution. Results run against much larger dataset may show improvements in the model's performance.

Unprepared accuracy is higher at 74% but low precision 29% and recall 8% with F_1 0.12. Kappa 0.03 indicates poor agreement. This result shows the peril of taking accuracy as a key indicator of model performance. The combination of unpromising results from the proxy features provides low precision and recall scores for unpreparedness. The model predicts 6.2% of EA as unprepared versus 32.5% by the expert. The model is outshone by the cognitive power of the expert to assemble proxy and additional features in classifying each EA. Unfavourable precision and recall scores again provide scant support for the unprepared hypothesis.

The models may show low recall and high precision because they only predict a positive when the sought-after feature is detected (i.e. the answer contains the trigger word). Answers that do not contain the trigger words increase the true negative count. The domain expert may class an answer as positive even without the appearance of trigger words by using other cues which are beyond the models' capabilities. Adherence to "bag-of-words" techniques where word order is discarded and hence the word's context is lost too may need the support of new methods which are better adapted to detecting context. The use of key phrases and world lists tuned to the domain of interest are more likely to improve model performance.

The results may have different meaning in different settings, one domain may require higher recall over precision and another vice versa. Similar domain dependent issues exist with the F_1 measure and Kappa. Although these results show less than impressive scores; a harmonic mean may be redundant when the balance of precision and recall is unimportant. Likewise, Kappa is a meaningful measure in the presence of high accuracy and class imbalance, it penalises the model for doing well due to class imbalance (Lantz; 2015). The consistently low Kappa values indicate a disagreement for each of the models.

The models are blunt instruments in comparison to the domain expert. Care is taken to preclude overfitting the annotated sample by the temptation to adjust the models to achieve optimum performance metrics. The objective is to select models that will generalise well in the wild.

A theory that may confound the models and related hypotheses is that an executive conveys a mixture of non-avoiding, certain and non-repetitive features even in a state being of unprepared. The domain expert and as (Bloomfield; 2012) suggests executives may have the ability to veil features and regulate their language. The coaching of executives could assist them to evade detection by these rudimental models. Executive may not be pressed to answer and hence may find it easier to command the discourse.

4.1 Limitations

Textual analysis in computational linguistics is a tricky area to navigate (Loughran and McDonald; 2016). The nuances in language whilst humanly simple to perceive are not easily recognisable by a computational model. The amount of data used in this exploration is small. Given that language is noisy, large amounts of data are required to identify the sought-after features. The more words, and hence features, that documents contain the more chance there is to discover the signal. Notwithstanding that non-contributing features are required to be removed.

The challenges to understand the language encountered and inferring meaning are many: loss of context, sarcasm, synonyms and polysemy, concept recognition, vagaries, cultural and personal inflections. The data input is unstructured and requires preprocessing prior to initial discovery. Computational techniques which attempt to glean knowledge from language rely on strict format and definition, clear logic and exactness (Fisher et al.; 2016). Existing words lists require expansion with words and phrases which are accepted and baselined in the target domain (Kearney and Liu; 2014).

This study is an initial investigation using a small annotated dataset with limited access to 1 domain expert. Conducting NLP in this domain which is littered with complexity and ambiguity is challenging. Finding techniques which are easy to demonstrate why they work and generalise well is demanding. The emphasis has been to adapt, combine and clearly show the performance of simple models. Considering the constraints, these preliminary results represent a reasonable first step with ample potential for future research.

5 Conclusion and Future Work

A initial sample of Executive answers (n=2457) is extracted from Q&A calls transcripts in the drinks industry. An exploration and observational study is conducted with the assistance of a domain expert. The objective is to investigate how proxy features of uncertainty, avoidance, repetition indicate executive preparedness in answering analyst questions. The domain expert participated in an annotation experiment using a subsample of EA (n=208).

Comparing the models' classification performances produced results which appear to disagree with the detection attained by the domain expert. Hence, the results lack the strength at feature level to support the hypotheses, accuracy is only just better than an 50:50 random guess together with disappointing F_1 and Kappa scores. The combined unprepared model in turn shows weak support for the overall preparedness hypothesis. Given the study's constraints the models are only evaluated against a small sample of EA. What might the results show in a larger EA corpus; comparing an executive's EA across time? How might the inter-company executive comparisons appear? Mixing this information with orthogonal data (e.g. market, political and legislative events) which impact the organisation may provide further intriguing insights. However, these simple models do provide information. Coaching executives to refrain from using the proxy features reduces the chance of executives leaking these linguistic cues, which may suggest support for the results of this exploration. In certain domains this information may provide a signal that merits a deeper inspection of the answer. The domain expert suggests that a tool which gauges communication features thus urges a closer human examination of the language and its meaning could be valuable to both the organisation and the investment

community. An application probing for executive ambiguities might assist investment managers to gain better insight of executive linguistic style and hence expose signals that hitherto may be overlooked. How well do executives articulate their operational strategies, instil audience confidence of their capability to execute these strategies? Could a high precision and low recall model be sufficient to identify outliers which are of interest to both the executive's company and the investor community?

Establishing meaning from language remains a broad area for future research and presents many challenges. Possible directions for future research in addition to those discussed include: Augmenting word lists with "executive tactic" trigger words (obtained from PR and financial communications companies) may provide new features to improve model performance; examining the discourse by inclusion of the analyst's question which initiated the EA and observing the interaction; application of supervised machine learning models; the use of video to provide additional features to analyse cues; analysis of languages other than English and moving beyond the current domain.

Acknowledgements

I would like to thank my supervisor Dr. Ralf Bierig for his valuable input, support and time afforded to me during this study. I am indebted to my domain expert Catherine James who so generously participated in the labelling experiment, explained investor relations and provided inspiration and insight. Thank you Dr. Simon Caton and Dr. Jason Roche for listening to my initial ideas and suggesting items for consideration. Finally, I thank my family for their unyielding support during the entire course and this research project.

References

Allee, K. D. and Deangelis, M. D. (2015). The structure of voluntary disclosure narratives: Evidence from tone dispersion, *Journal of Accounting Research* **53**(2): 241–274.

Allport, G. W. (1961). *Pattern and growth in personality*, Holt, Rinehart and Winston, New York.

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I. and Edu, J. B. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research* **3**: 993–1022.

Bloomfield, R. (2012). Discussion of detecting deceptive discussions in conference calls, *Journal of Accounting Research* **50**(2): 541–552.

Bozanic, Z. and Thevenot, M. (2015). Qualitative Disclosure and Changes in Sell-Side Financial Analysts' Information Environment, *Contemporary Accounting Research* **32**(4): 1595–1616.

Brockman, P., Cicon, J. E., Li, X. and Price, S. M. (2017). Words versus deeds: Evidence from post-call manager trades, *Financial Management* pp. n/a–n/a.

Brockman, P., Li, X. and Price, S. M. (2015). Differences in conference call tones: managers vs. analysts, *Financial Analysts Journal* **71**(4): 24.

Burgoon, J., Mayew, W. J., Giboney, J. S., Elkins, A. C., Moffitt, K., Dorn, B., Byrd, M. and Spitzley, L. (2016). Which spoken language markers identify deception in high-stakes settings? evidence from earnings conference calls, *Journal of Language and Social Psychology* **35**(2): 123–157.

Chen, H., De, P., Hu, Y. and Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media, *Review of Financial Studies* **27**(5): 1367–1403.

Davis, A. K., Ge, W., Matsumoto, D. and Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls, *Review of Accounting Studies* **20**(2): 639–673.

Doran, J. S., Peterson, D. R. and Price, S. M. (2012). Earnings Conference Call Content and Stock Price: The Case of REITs, **45**(2): 402–434.

Duran, N. D., Hall, C., Mccarthy, P. M., Mcnamara, D. S. and Duran, N. (2010). The linguistic correlates of conversational deception: Comparing natural language processing technologies, *Applied Psycholinguistics* **31**: 439–462.

Fisher, I. E., Garnsey, M. R. and Hughes, M. E. (2016). Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research, *Intelligent Systems in Accounting, Finance and Management* **23**(3): 157–214.

Ghiassi, M., Skinner, J. and Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network, *Expert Systems with Applications* **40**(16): 6266–6282.

Gunsch, M., Brownlow, S., Haynes, S. and Mabe, Z. (2000). Differential linguistic content of various forms of political advertising., *Journal of Broadcasting and Electronic Media* **44**(1): 27–42.

Guo, L., Shi, F. and Tu, J. (2017). Textual analysis and machine leaning: Crack unstructured data in finance and accounting, *The Journal of Finance and Data Science* pp. –. **URL:** *http://www.sciencedirect.com/science/article/pii/S2405918816300496*

Henry, E. (2008). Are Investors Influenced By How Earnings Press Releases Are Written?, *Journal of Business Communication* **45**(4): 363–407.

Hobson, J. L., Mayew, W. J. and Venkatachalam, M. (2012). Analyzing speech to detect financial misreporting, *Journal of Accounting Research* **50**(2): 349–392.

Hollander, S., Pronk, M. and Roelofsen, E. (2010). Does silence speak? an empirical analysis of disclosure choices during conference calls, *Journal of Accounting Research* **48**(3): 531–563.

Ittoo, A., Nguyen, L. M. and van den Bosch, A. (2016). Text analytics in industry: Challenges, desiderata and trends, *Computers in Industry* **78**: 96 – 107.

Kearney, C. and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models, *International Review of Financial Analysis* **33**: 171–185.

Kumar, B. S. and Ravi, V. (2016). A survey of the applications of text mining in financial domain, *Knowledge-Based Systems* **114**: 128 – 147.

Lantz, B. (2015). *Machine Learning with R*, 2nd edn, Birmingham: Packt.

Larcker, D. F. and Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls, *Journal of Accounting Research* **50**(2): 495–540.

Lee, J. (2016). Can Investors Detect Managers' Lack of Spontaneity? Adherence to Predetermined Scripts during Earnings Conference Calls, *The Accounting Review* **91**(1): 229–250.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence, *Journal of Accounting and Economics* **45**(2-3): 221–247.

Li, F. (2010). Textual analysis of corporate disclosures: A survey of the literature, *Journal of Accounting Literature* **29**: 143.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks., *Journal of Finance* **66**(1): 35 – 65.

Loughran, T. and McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey, *Journal of Accounting Research* **54**(4): 1187–1230.

Manning, C. D. (2015). Computational linguistics and deep learning., *Computational Linguistics* **41**(4): 701 – 707.

Manning, C. D., Raghavan, P. and Schutze, H. (2008). *An introduction to information retrieval.*, Cambridge: Cambridge University Press.

Matsumoto, D., Pronk, M. and Roelofsen, E. (2011). What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions, *Accounting Review* **86**(4): 1383–1414.

Mayew, W. J. (2008). Evidence of management discrimination among analysts during earnings conference calls., *Journal of Accounting Research* **46**(3): 627 – 659.

Mayew, W. J. and Venkatachalam, M. (2012). The power of voice: Managerial affective states and future firm performance., *Journal of Finance* **67**(1): 1 – 44.

McDonald, L. (1990). 'No Comment' The Art of Avoiding the Question, *ETC: A Review of General Semantics* **47**(1): 8–14.

Newman, M. L., Pennebaker, J. W., Berry, D. S. and Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles, *Personality and Social Psychology Bulletin* **29**(5): 665–675.

Palmieri, R., Rocci, A. and Kudrautsava, N. (2015). Argumentation in earnings conference calls. corporate standpoints and analysts challenges, *Studies in Communication Sciences* **15**(1): 120 – 132.

Pennebaker, J. W. and King, L. A. (1999). Linguistic styles: Language use as an individual difference., *Journal of Personality and Social Psychology* **77**(6): 1296 – 1312.

Price, S. M., Doran, J. S., Peterson, D. R. and Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone, *Journal of Banking and Finance* **36**(4): 992–1011.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50.

Schmidt, S., Schnitzer, S. and Rensing, C. (2016). Text classification based filters for a domain-specific search engine, *Computers in Industry* **78**: 70–79.

Strang, K. D. and Sun, Z. (2016). Analyzing Relationships in Terrorism Big Data Using Hadoop and Statistics, *Journal of Computer Information Systems* **4417**(October 2016): 1–9.

Sun, S., Luo, C. and Chen, J. (2017). A review of natural language processing techniques for opinion mining systems, *Information Fusion* **36**: 10 – 25.

Szarvas, G., Farkas, R., Vincze, V. and Mra, G. (2012). Cross-genre and cross-domain detection of semantic uncertainty., *Computational Linguistics* **38**(2): 1 – 58.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods, *Journal of Language and Social Psychology* **29**(1): 24–54.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market., *Journal of Finance* **62**(3): 1139 – 1168.

Thomas, A. K., Chen, C., Gordon, L. T. and Tenbrink, T. (2015). Choose Your Words Wisely: What Verbal Hesitation Indicates About Eyewitness Accuracy, *Applied Cognitive Psychology* **29**(5): 735–741.

Uysal, A. K. and Gunal, S. (2014). The impact of preprocessing on text classification, **50**(1): 104–112.

Vrij, A. (2008). *Detecting Lies and Deceit Second Edition*, John Wiley and Sons, Chichester, UK.

A Configuration Manual

This appendix describes the configuration and process to follow to recreate the study's environment, preparation of data, creating the sample spreadsheet of executive answers for annotation, executing the models to evaluate their performance in comparison to the domain expert annotated dataset.

A.1 Environment and Set up

The environment on which the research is conducted is a laptop running Windows 10 64-bit with an intel core i5 2.3GHz, 32GB RAM and 1TB SSD disk.

A.1.1 Installation Software Components

Anaconda[1] is selected as it offers a comprehensive easy to install platform for conducting Python based project development. Install Anaconda[2] for Python 2.7 at download Anaconda. Installing into a new environment is probably the cleanest way to start unless a python 2.7.12 environment is already in place. This configuration has not been tested with Python 3 and may produce unexpected results. Once installed, the following setup information should be similar to that in Figure 1 when typing conda info from the command prompt.

```
Current conda install:

               platform : win-64
          conda version : 4.3.21
     conda is private : False
    conda-env version : 4.3.21
  conda-build version : 2.0.2
       python version : 2.7.12.final.0
     requests version : 2.12.4
     root environment : C:\Users\tomd\Anaconda2  (writable)
  default environment : C:\Users\tomd\Anaconda2\envs\pda
     envs directories : C:\Users\tomd\Anaconda2\envs
                        C:\Users\tomd\AppData\Local\conda\conda\envs
                        C:\Users\tomd\.conda\envs
        package cache : C:\Users\tomd\Anaconda2\pkgs
                        C:\Users\tomd\AppData\Local\conda\conda\pkgs
```

Figure 1: Anaconda Configuration

Required Python Packages:

- NLTK 3.2.1
- sci-kitlearn 0.17.1
- numpy 1.11.1

[1]https://www.continuum.io [Accessed 16 August 2017]
[2]https://www.continuum.io/downloads [Accessed 16 August 2017]

- pandas 0.18.1
- scipy 1.18.1
- Genism 2.2.0[3]

Install a Python IDE, the project code is run from the IDE. The choice here is Spyder 3.1.4 and can be installed following instructions at download Spyder[4]. Microsoft Excel is required or any application capable of working with xlsx format. A text editor (e.g. Sublime[5]) is used for CSV, JSON and world list editing.

If running the Annotation Sample experiment, the following additional software is required.

- Pymongo 3.3.1
- MongoDB 3.2.10

A standard MongoDB setup is used. Follow install instructions at download MongoDB[6]. Running mongod process from the command prompt should start MongoDB and display output similar to Figure 2.

```
C:\Users\tomd>mongod
2017-07-02T10:36:16.254+0100 I CONTROL  [initandlisten] MongoDB starting : pid=7096 port=27017 dbpath=C:\data\db\ 64-bit
 host=DESKTOP-OAE8JQV
2017-07-02T10:36:16.321+0100 I CONTROL  [initandlisten] targetMinOS: Windows 7/Windows Server 2008 R2
2017-07-02T10:36:16.383+0100 I CONTROL  [initandlisten] db version v3.2.10
2017-07-02T10:36:16.388+0100 I CONTROL  [initandlisten] git version: 79d9b3ab5ce20f51c272b4411202710a082d0317
2017-07-02T10:36:16.405+0100 I CONTROL  [initandlisten] OpenSSL version: OpenSSL 1.0.1t-fips  3 May 2016
2017-07-02T10:36:16.428+0100 I CONTROL  [initandlisten] allocator: tcmalloc
2017-07-02T10:36:16.453+0100 I CONTROL  [initandlisten] modules: none
2017-07-02T10:36:16.472+0100 I CONTROL  [initandlisten] build environment:
2017-07-02T10:36:16.497+0100 I CONTROL  [initandlisten]     distmod: 2008plus-ssl
2017-07-02T10:36:16.500+0100 I CONTROL  [initandlisten]     distarch: x86_64
2017-07-02T10:36:16.518+0100 I CONTROL  [initandlisten]     target_arch: x86_64
2017-07-02T10:36:16.528+0100 I CONTROL  [initandlisten] options: {}
2017-07-02T10:36:16.580+0100 I -        [initandlisten] Detected data files in C:\data\db\ created by the 'wiredTiger' s
torage engine, so setting the active storage engine to 'wiredTiger'.
2017-07-02T10:36:16.603+0100 I STORAGE  [initandlisten] wiredtiger_open config: create,cache_size=4G,session_max=20000,e
viction=(threads_max=4),config_base=false,statistics=(fast),log=(enabled=true,archive=true,path=journal,compressor=snapp
y),file_manager=(close_idle_time=100000),checkpoint=(wait=60,log_size=2GB),statistics_log=(wait=0),
2017-07-02T10:36:21.170+0100 I NETWORK  [HostnameCanonicalizationWorker] Starting hostname canonicalization worker
2017-07-02T10:36:21.182+0100 I FTDC     [initandlisten] Initializing full-time diagnostic data capture with directory 'C
:/data/db/diagnostic.data'
2017-07-02T10:36:21.211+0100 I NETWORK  [initandlisten] waiting for connections on port 27017
2017-07-02T10:37:22.955+0100 I NETWORK  [initandlisten] connection accepted from 127.0.0.1:1888 #1 (1 connection now ope
n)
```

Figure 2: MongoDB Configuration

A.2 Source Code and Data

Once the environment is set up, create the following folders changing the names to suit:

- Source code e.g. c:\eaCode\
- Source data e.g. c:\eaData\
- Output for each executive answer e.g. c:\data\textout\execEach\

[3]http://radimrehurek.com/gensim/install.html [Accessed 16 August 2017]
[4]https://github.com/spyder-ide/spyder [Accessed 16 August 2017]
[5]https://www.sublimetext.com/3 [Accessed 16 August 2017]
[6]https://docs.mongodb.com/ [Accessed 16 August 2017]

The source code can be downloaded from github at: prepared executive source code[7] and copied to the local source code folder. The raw Q&A call transcripts and word lists can be downloaded from: prepared executive source data code[8] and copied to the source data folder.

A.3 Processing Executive Answers

This is a 2 step process using the raw Q&A dataset to create the executive and analyst sequence file and then using that file to create the executive answers extract file. Figure 3 illustrates the process. Begin by unzipping the downloaded raw Q&A and word lists into the data source folder.

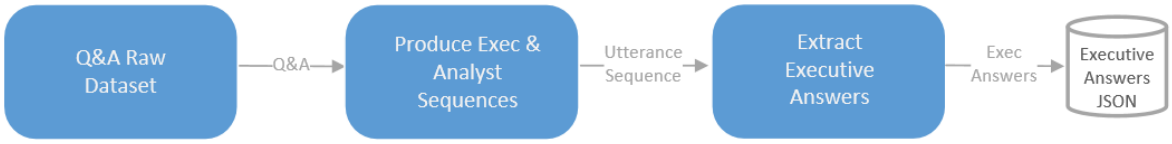

Figure 3: Extracting Executive Answers

A.3.1 Create Executive and Analyst Sequences

Open unwinder_exec_anal_v1.0.py in Spyder and amend the path and input file name to read the appropriate conference call transcript file (e.g. transcripts_may31_TAP_Q1_17.json). The script processes one call transcript at a time but can be amended to read the input files from a folder if desired. Running the script produces the Executive and Analyst sequences JSON file, a sample of part of an input file e.g. transcripts_may31_TAP_Q1_17.json is shown below:


```
[{''entry'': {''title": ''Q1 2017 Earnings Conference
Call'', ''company": ''Molson Coors Brewing Co'', ''exec": [[''Tracey
Joubert'', ''Global CFO", ''Molson Coors Brewing Co''], [''Mark
Hunter", ''President & CEO'', ''Molson Coors Brewing Co'']],
''analysts'': [[''Vivien Azer'', ''Cowen and Company'']], [''Andrea
Teixeira'', ''JPMorgan'']], [''Judy Hong'',
''Goldman Sachs'']], ''questions'': [''Vivien Azer '', ''So, Mark,
can I just follow-up on your commentary around the improvement that
you've seen in the U.S. post soft January and February, can you
quantify the rate of improvement?'', ''Mark Hunter'', ''Thanks,
Vivien. I can't quantify the rate improvement, but I mean if you
look at the scanner data, then we've...'']}}]
```


[7]https://github.com/datamin3r/QandA/tree/master/research-project/code [Accessed 16 August 2017]

[8]https://github.com/datamin3r/QandA/tree/master/research-project/code [Accessed 16 August 2017]

The script outputs the executive and analyst sequences JSON file, a sample is given below from transcript_may31_seq_resulttranscripts_may31_TAP_Q1_17.json


```
[{''entry'': {''executivesNames'': [[''Tracey Joubert'',
''Global CFO'', ''Molson Coors Brewing Co'']],
...

''posexec'': [{''pos'': [91,117,143,171,175],''exec'': ''Tracey
Joubert''},
...
''execs'': [{''answer'': ''Yes. So, Andrew, what we did say is that
our cash tax benefit would be around \$275 million on average over
15 years. We did say that it would be frontend loaded...'',
''sequence'': 92, ''exec'': ''Tracey Joubert'' }]}}]
```


A.3.2 Create Executive Answers

Open extractExecUttersEach_v1.0.py in Spyder and amend the path and input file name of the call to select (e.g. transcript_may31_seq_resulttranscripts_may31_TAP_Q1_17.json) and set the executive's name to extract their answers. The executives name can be found by inspecting the JSON file. The script processes one executive at a time to enable the selection of the desired executives. Running the script will output the Executive Answers JSON file, a sample is provided below.


```
[{''execResp'': [''The thing I would add to a, market is that
we haven't seen impact on our premium brands effect we actually
increase share meaningfully in the fourth quarter and December
was actually one of our best year months in a very long time.
In economy strategy went to back into the year. In fact we grew
segment share with Miller Lite and Coors Light for 20
consecutive months. I haven't seen that impact that Stephen
refers to and is certainly not a strategy. '', ''So we have
plans in place. We're making costs possible because it's very
difficult very early to predict andcomment on any kind of
changes that will come from tax reform that the administration
is looking at. We have plans, we plan on if something changes
we will make sure that they we can change direction.''],
''execName'': ''Tracey Joubert'', ''execRespCount'':18}]
```


A.4 Executive Sample Labelling Experiment

This section describes the process shown in Figure 4 to create the sample of executive answers in a spreadsheet which the domain expert annotates.

Check the executive answers files from the previous step are in the folder created as c:\eaData\textoutexecEach\. Open putexecAnswers.py in Spyder and amend the path and to select all the executive answer files in that folder. Start MongoDB (open a command prompt and type mongod, open a new command window and to start the mongo shell type mongo). Run the python script putexecAnswers.py in Spyder and then

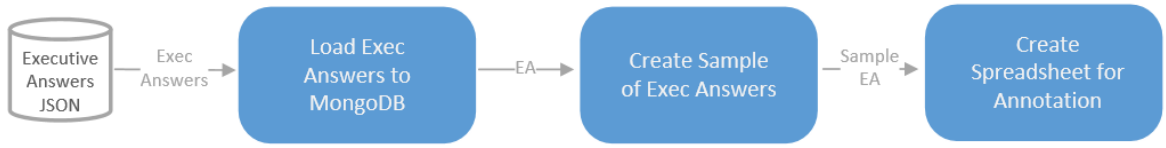

Figure 4: Process to Create Sample Spreadsheet for Labelling Experiment

check that the executive answers are loaded into MongoDB as shown in the command sample in Figure 5.

```
> use execsamp
switched to db execsamp
> db.exectrans7.find().count()
2457
> db.exectrans7.findOne()
{
        "_id" : ObjectId("5958ed1191f6c800d8f8912c"),
        "answer" : "May be Gilles you want to take the Border Adjustment Tax and I'll take the other two?",
        "conCallAns" : "31_Pernod_Q2_17_0",
        "execName" : "Alexandre Ricard"
}
>
```

Figure 5: Checking Executive Answers exist in MongoDB

In Spyder open execAnswers.py and check the path matches the folder created earlier. Run the script which produces the 208 executive answers sample csv file shown in Figure 6. This file is copied and updated with the annotated results [A.5] prior to model execution.

	A	B	C	D	E
1	conCallAn	execAnswer	execName		
2	ay29const	And you can cal	Rob Sands		
3	y31_HEIN'	Yes. I want to b	Jean-Francois van Boxmeer		
4	_the_Pres	What we're see	John Kennedy		
5	_BrownF.I	It certainly was	Paul Varga		
6	y31_HEIN'	Competitive dyr	Jean-Francois van Boxmeer		

Figure 6: Executive Answers CSV

Create a spreadsheet from this csv file keeping the executive answers and add the new columns displayed in Figure 7 (Generate the AnsNo cells data using this formula =ROW(A2)-1). Save this file as the master sample of unannotated executive answers.

	A	B	C	D	E	F
1	AnsNo	execAnswer	Uncertainty	Avoidance	Repetition	Unprepared
2	1	And you can call it leading or not leading. We are sort of I would say marching to our own beat as opposed to too worried about what some of our competitors are doing overall. And that's sort of up to them. And we will do what we think the market will bear for our brands and what's good for the health of our brands.				
3	2	Yes. I want to be crystal clear on that management is appointed by the board, and has the full competence of the board to carry these forward. That's one. Secondly, SEBI has issued a executive order to this mister Chairman of the Board out of his function. It's not the will of the shareholders to do it nor on board of directors, it is just a court order that have to be executed. It's for Dr. Mallya to react on that if he wants to have a stay on that position.				
	3	What we're seeing is it's positive for the brand equity. We also have figured out a lot of the principles how to merchandise, which makes sure that as you launch the new product, you use it				

Figure 7: Executive Answers CSV showing new columns added

A.5 Model Execution

The section describes the execution of the models using 2 python scripts and the classified executive answers and the corresponding unannotated executive answers. The workflow is shown in Figure 8.

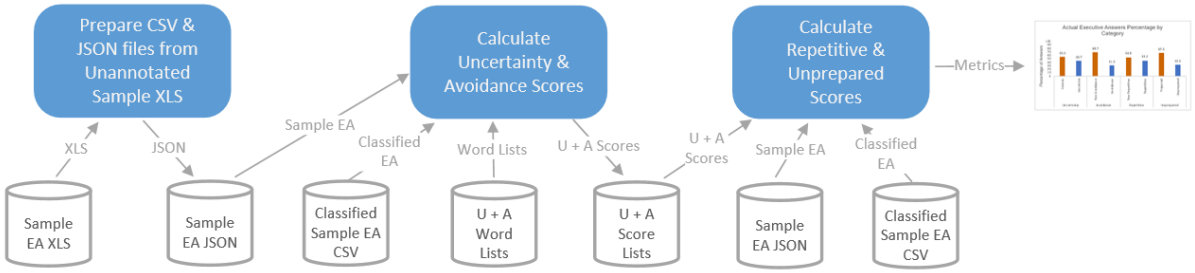

Figure 8: Model Execution Workflow

The annotated spreadsheet having been completed by the domain expert is merged manually with a copy of the executive answers csv file [A.4] to create a new classifiedExecAnswers_197_v0.1.csv file, a sample is shown in Figure 9.

	A	B	C	D	E	F	G	H
1	conCallAn	execAnsw	execName	AnsNo	Uncertaint	Avoidance	Repetition	Unprepared
2	ay29const	And you c	Rob Sands	1	-1	-1	1	-1
3	y31_HEIN	Yes. I wan	Jean-Franc	2	1	1	1	0
4	_the_Pres	What we'r	John Kenn	3	1	0	1	0
5	_BrownF.I	It certainly	Paul Varga	4	-1	-1	1	-1
6	y31_HEIN	Competiti	Jean-Franc	5	1	0	1	0
7	_may31_T	We've alsc	Michael Cl	6	-1	0	0	0
8	_BrownF.I	I know at	Paul Varga	7	-1	-1	-1	-1
9	may31_TA	Yeah, I wa	Gavin Hatt	8	-1	-1	-1	-1
10	_may31_T	The way t	Michael Cl	9	-1	-1	-1	-1

Figure 9: Classified Executive Answers CSV

Create a JSON format file of the corresponding unannotated answers using the python

script annotatedExecAnsJsonConverter.py which takes an unannotated saved copy of the master spreadsheet (e.g.ExecUnannotated_for_json_input.csv) as input and produces UnannotatedExecAnswers.json. 197 of the executive answer are used following the removal of 11 answers that were continuity responses and 1 question. Ensure that both input files contain the same number of answers. The word list files used by the models should reside in the the data folder having already been downloaded. The Loughran and McDonald uncertainty word list is freely available on University of Notre Dame[9]. However, the LIWC word lists[10] are only available for non commercial research use by schools and univerties, the lists are not to be shared outside of the research lab.

A.5.1 Calculating Uncertainty and Avoidance

Using the 2 files created above as input (i.e. UnannotatedExecAnswers_197.json and classifiedExecAnswers_197_v0.1.csv) open annotatedExecAnsPerformance.py in Spyder and run the script which produces the confusion matrix and performance measures. Also check the path to ensure that the pickle files produced are written to the desired folder. The following figures show the output from the study's data. The counts of annotated executive answers by feature are given in Figure 10. The confusion matrix and model

```
Annotated Answers
-----------------
Uncertainty
Certain      99
Uncertain    80
Neutral      18
Name: Uncertainty Measures, dtype: int64

Avoidance
Non Avoidance    112
Avoidance         51
Neutral           34
Name: Avoidance Measures, dtype: int64

Repetition
Non Repetition    80
Repetition        66
Neutral           51
Name: Repetition Measures, dtype: int64

Unprepared
Prepared      104
Unprepared     50
Neutral        43
Name: Unprepared Measures, dtype: int64
```

Figure 10: Counts of Annotated Executive Answers by Feature

performance metrics for Uncertainty in the sample of executive answers are shown in Figure 11. The confusion matrix and model performance metrics for Avoidance in the

[9]http://sraf.nd.edu/textual-analysis/resources/ [Accessed 16 August 2017]
[10]http://liwc.wpengine.com/compare-dictionaries/ [Accessed 16 August 2017]

```
------------------
Performance Metrics
------------------

Uncertainty

Pred Certainty count   =  120
Pred Uncertainty count =   59


 Uncertainty Confusion Matrix

[[66 33]
 [54 26]]

Accuracy        0.51
Precision       0.44
Recall          0.33
F1              0.37
Kappa           -0.01
```

Figure 11: Uncertainty Model Performance

sample of executive answers are shown in Figure 12. The confusion matrix and model

```
Avoidance

Pred Non Avoidance count =   155
Pred Avoidance count     =   8


 Avoidance Confusion Matrix

[[109   3]
 [ 46   5]]

Accuracy        0.70
Precision       0.62
Recall          0.10
F1              0.17
Kappa           0.09
```

Figure 12: Avoidance Model Performance

performance metrics for Repetition in the sample of executive answers are shown in Figure 13. The confusion matrix and model performance metrics for Unprepared in the sample of executive answers are shown in Figure 14.

```
Performance Metrics
-------------------
Pred Non Repetition count =  113
Pred Repetition count     =   33


 Repetition Confusion Matrix

[[64 16]
 [49 17]]

Accuracy        0.55
Precision       0.52
Recall          0.26
F1              0.34
Kappa           0.06
```

Figure 13: Repetition Model Performance

```
 Unprepared

 Pred Unprepared count =  7
 Pred prepared count  =   106


  Unprepared Confusion Matrix

 [[82  5]
  [24  2]]

 Accuracy        0.74
 Precision       0.29
 Recall          0.08
 F1              0.12
 Kappa           0.03
```

Figure 14: Unprepared Model Performance