西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

# 2018 届
# 本科毕业论文（设计）

论文题目：Analyst Characteristics, Textual Information and Prediction Accuracy: Evidence from China Financial Market

学生姓名：闫嘉文

所在学院：经济与管理研究院

专　　业：经济学(经济与管理国际化创新人才班)

学　　号：41415206

指导教师：杜茜茜

成　　绩：

2018 年 04 月

# 西南财经大学

# 本科毕业论文原创性及知识产权声明

本人郑重声明：所呈交的毕业论文是本人在导师的指导下取得的成果，论文写作严格遵循学术规范。对本论文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。因本毕业论文引起的法律结果完全由本人承担。

本毕业论文成果归西南财经大学所有。

特此声明

毕业论文作者签名：

作者专业：

作者学号：

年　　月　　日

# Abstract

**This paper examines the connection between analyst characteristics and their performance. Employing a Naïve Bayes algorithm and key facial landmarks detection approach on analyst profiles and text in analyst reports, we extract textual opinion from analyst reports and analyst personality traits from their standard profile images. We find that both intrinsic and extrinsic analyst characteristics are correlated with the qualitative textual descriptions in analyst reports, but not with quantitative EPS forecasts; and certain welcomed attributes reduce analysts' predicting errors. We also find that analysts with higher attractiveness are more likely to have favorable career outcomes.**

**Keywords: analyst characteristics; textual analysis; personality traits; analyst performance**

# 摘要

本文研究了中国证券市场卖方分析师个人特征与其业绩表现之间的潜在关系。通过朴素贝叶斯机器学习法和人像特征点识别法，我们分析并提取了研究报告的文本特征，并从公开证件照中提取得到分析师性格特征。通过研究我们发现分析师的内在性格特征和外在固定特征与分析师文本信息特征显著相关，但与预测每股盈余等数量信息没有显著关系；但拥有受欢迎特征分析师的预测更加准确。我们还发现，拥有年轻漂亮等受欢迎特质的卖方分析师更容易在职业生涯中获得成功，成为明星分析师。

关键字：分析师特征；文本分析；人格特质；分析师业绩

# Analyst Characteristics, Textual Information and Prediction Accuracy: Evidence from China Financial Market

## Contents

# 1. Introduction

Sell-side analysts, who collecting, analyzing, and delivering information to both institutional and individual investors, act at a pivotal position in the financial market. Consequently, an avalanche of literatures has been devoted to investigating forecast behavior, such as analysts' quantitative measures' effectiveness (Twedt and Rees, 2012), the information content of analyst reports (Huang et al., 2014), and analysts' performance with extrinsic experiences (Bradley et al., 2017). Despite an observable number of "innate gift" theory of analyst occupation, there is little empirical evidence linking analyst characteristics and their overall performance due to the difficulties in measuring intrinsic characteristics. Thus, the influence of analysts' intrinsic characteristics on forecasting is underdeveloped. Given the importance of such "innate talent" in sociology and psychology (Bainbridge, Isola, and Oliva, 2013), we attempt to investigate analysts and their performance from a novel perspective using comprehensive Chinese analyst data.

In this study, we employ key fiducial points detection approach to extract three key personality traits from analysts' standard profile images and employ Naïve Bayes algorithm to extract textual sentiment opinion from huge number analyst reports. By linking analysts' characteristics, both intrinsic and extrinsic with their working product – analyst reports, we are able to perform a comprehensive research answering what factors could influence analysts' performance. Specifically, we explore three substantial sub-related topics: (1) What analysts' characteristics could direct features of their textual reports? (2) What analysts' characteristics could shad influence on their prediction accuracy? (3) And when shifting perspective to long-term career path, what analysts' characteristics drive them to have successful careers in analysts' universe?

To investigate analyst characteristics on textual reports, prediction accuracies and further career outcome, we collect 117,100 registered analyst standard profile images from SAC, the Security Association of China, with pre-wrapped Python modules facilitating us to extract 68 key fiducial landmarks. Deriving 62 facial attributes revised from Vernon et al. (2014)'s practice by extracted 68 landmarks allow us to get three personality traits – Approachability, Youthfulness and Attractiveness, and Dominance

scores. We also employ a pre-trained Naïve Bayes classifier to extract textual opinion of 1,068,679 analyst reports for China listed companies from 2006 to 2016. Specifically, each sentence is classified into either positive, neutral, or negative, and then be aggregated together to derive the overall opinion of a given analyst report. Our validation tests prove that this machine learning approach currently outperforms traditional dictionary methods, word vectors approach, and sentiment analysis API available online for China financial market data.

We first describe the general profile image as the beginning of our empirical analysis. A 3-by-3 figure is illustrated to exhibit low-mid-high examples of three intrinsic personality traits. Textual analyst reports are also described. An ordinary analyst report would comprise 20 sentences, 72% of them being positive, 17% of them being neutral and remaining 11% being negative. A figure illustrating temporal fluctuations among analyst report length, analyst report opinion, and a number of reports released from 2006 to 2016 is made to vividly depict co-movement with Shanghai Stock Composite Index's boom and bust in mid-2007 and mid-2015.

In our first empirical model, we document the influence of analyst intrinsic and extrinsic characteristics on their soft information report opinion but not on hard recommendation change and forecast EPS. Specifically, we find that post site-visitation reports' opinion level is 7.88% higher, and that personality traits significantly alter report opinion level – approachability negatively adjusting opinion level; youthfulness and attractiveness and dominance positively adjusting opinion level. These results indicate information content of analyst reports is not only guided by firm and industry fundamentals but also augmented by analysts' characteristics.

Additionally, to better understand analyst characteristics' influence on different perspectives of textual contents, we construct six textual related variables. Existing study (Huang et al., 2014) examined the validness and effectiveness of those textual characteristics on U.S. market. We find that dominant and star analysts incline to use more confident and assertive words, but opposite for female analysts. Financially related discussions are less proportionated when analysts are junior and when conducted site-visitation recently, but reversely for female and high educated analysts. Other analyst characteristics also separately have impacts on different textual characteristics.

We next explore the relationship between analyst characteristics and their corresponding prediction accuracy. After controlling for textual opinion, firm heterogeneity and industry and year fixed effects, we contend that analysts' absolute forecast errors (*AFE*) are lower when analysts are female, when analysts obtained high education degrees, when analysts are *not* Star analysts, and when affiliated trading house size is small. By employing Difference-in-Difference (DID) approach, we find that one standard deviation increases of textual opinion in treatment (site-visiting) groups, ironically, increases analysts forecast errors by 5 base points.

Given previously proved evidence of high relation between analyst characteristics and prediction accuracy, we finally examine whether welcomed characteristics could provide analysts with favorable career outcomes. We contend that, on average, young and attractive analysts are 7.50% more probable to be "blossomers"; and in this seemingly male dominating world, female analysts encounter greater obstacles to become career successful, with 22.65% fewer odds being elected as Star analysts. We also find that analysts' likelihood to be elected as Star analyst is not determined by year average performance, but by several individual outstanding reports.

Overall, our paper highlights the importance of analyst characteristics to textual report quality, prediction accuracy, and career outcome by comprehensively analyzing analysts' both intrinsic and extrinsic characteristics. Our study also fills the blank in the literature of comprehensive analyst report textual analysis in China financial market by extracting and exploring large-sample analyst reports and by combining textual reports with personality traits.

Our study provides several distinct insights for analyst universe than previous workings. First, we take the first step to use standard profile images to extract analyst intrinsic characteristics. Second, we document the effectiveness of Naïve Bayes Algorithm for Chinese textual content in China financial market. Third, we conduct both report level and market level analysis to evaluate the effectiveness of analyst characteristics, and illustrate that analysts with certain exceptional characteristics tend to have higher report quality, lower forecasting error, and more favorable career outcome. Last, our large analyst reports dataset and profile image set provides greater generalizability of our results.

The remaining of this paper expands as follows. Section II discusses related prior studies; section III develops the hypothesis; section IV introduces personality traits extraction approach and Naïve Bayes textual classification algorithm; section V describes sample selection and major variables; section VI presents our empirical results; and section VII raises our final conclusions.

# 2. Prior Studies

Researches about understanding text information and corresponding market reactions have surged in recent years with great progress in natural language processing (NLP). Studies examined textual content informativeness in a variety of context in financial universe, including financial related public news (Piotroskia et al., 2016), analyst reports (Huang et al., 2014), and management discussion and analysis (MD&A) (Li, 2010). Interrelations between textual analysts' reports and corporate earnings announcements are also examined to uncover analysts' information reinterpretation role and information discovery role in the financial market (Chen et al., 2009; Huang et al., 2016).

Machine learning classification approach has been proved to be one of the most successful approaches to dealing with natural language classification problem. With proper predefined training algorithms and hyper parameters, a classifier could be trained to identify and, further, to learn informative features in text, based on statistical frequency of appearing; and then be utilized to perform domain related context classification tasks. This statistical approach works both faster and better than the traditional bag of words (BOW), or dictionary, method in measuring the textual tone (Li, 2010; Li, 2011; Huang, 2014; Piotroski et al., 2016). By manually classifying 10,000 sentences' category and training with Naïve Bayes algorithm, a "naive" but effective algorithm in natural language processing (NLP), Huang et al. (2014) obtained a classifier with 75% robust accuracy and documented additional information content role of analyst reports with 2% two-day cumulative abnormal returns, CARs, for S&P 500 firms by longing the top quintile and shorting the bottom quintile simultaneously. Distinctive as China financial

market may be from western developed financial markets, similar textual additional information role is found in listed company's notices by Yan, Wang, and Kang (2017).

Major existing literatures explored and documented the influence of analysts' extrinsic characteristics, such as industry experiences, social connections, affiliated trading houses etc., on their prediction accuracy, career path, corresponding financial market reaction etc. (Bradley et al., 2017; Fang and Huang, 2017). Bradley et al. (2017), employing past working experience as a proxy of ones' social connection, and Fang and Huang (2017), directed measuring analyst-executive connection using data from BoardEx database, both proved that socially well-connected analysts occupy a pivotal position in analysts' networks and, consequently, have a greater likelihood to enjoy a favorable career path – becoming "Star" analysts.

By contrast, few studies focus on analysts' intrinsic, or "innate" characteristics, despite its potentially wild applications in fields of politics (Joo et al., 2017) and psychology (Bainbridge, Isola and Oliva, 2013). Unlike extrinsic qualities, intrinsic characteristics capture individual "soft" qualities, say, personality traits, may reveal their working behaviors and corresponding consequences. Existing studies measure intrinsic quality from rather simple and facet perspective. Jia et al. (2014) examined the relationship between male executives' masculinity and financial misreporting by measuring their facial height-to-width ratio (*fWHR*), given the potential relationship of facial masculinity and masculine working behaviors (aggression, risk-seeking, etc.) in males. He et al. (2016) exhibited that wall-street analysts with strong personal achievement drives tend to have a higher *fWHR* ratio. Admittedly, simple as measure as *fWHR*, though effective in certain scenarios, could be subjective and unreliable, for the existence hundreds number of key fiducial points and corresponding dozens of facial attributes for a given face (Le et al., 2012; Vernon et al., 2014)

Yet, with progress of psychology and computer science, identifiable personality traits are reasonably developed. Vernon et al. (2014) modelled 65 facial attributes by manually localizing 187 key facial landmarks and linearly relating those attributes to three personality traits – approachability, youthfulness and attractiveness, and dominance, explaining 58% of variances despite environment variations in profile images. Pre-labeled datasets, iBUG dataset (Sagonas et al., 2016) and Helen dataset (Le et al., 2012), are publicly available for training a decent facial landmark detector portraying 68 and 194

key facial feature points, or landmarks, respectively, automating this personality traits scoring process. By linearly combining facial attributes computed by model detected key facial landmarks, following Vernon et al. (2014), one's personality traits scores could be developed in short. Another personality traits identification procedure is realized by transfer learning a trained convoluted neural network (CNN) model, since training from scratch with little data may easily encounter the over-fitting problem (Li, Johnson, and Yeung, 2017). Often, the final fully connected layers of a pre-trained CNN model, filtering and capturing features of images, are toned to capture and to predict one's personality traits. This approach works only slightly better than traditional feature evaluation and principal component analysis (PCA) approach, improving accuracy by about 5% for 20 listed traits and may be easily over-fitted with small samples used by Zhang et al. (2017).

# 3. Hypotheses Development

## 3.1 Revealing Analysts' Characteristics from Textual Analyst Reports

Writing high quality analyst report is often regarded as one of the most important working components of sell-side analysts besides conducting in-field site visit, carrying out thorough industry and firm researches, and delivering roadshows and promoting reports in a variety of context[1]. As a final product synthesizing their professional knowledge, information acquired from site visits and researches, analyst reports often represent analysts' productiveness, competence and potential. A typical analyst report will give both "hard" quantitative measures, such as recommendation level, recommendation revision, forecasted earnings per share (EPS), target prices, etc., as well as "soft" qualitative textual descriptions, such as events analysis, forecasted operating

[1] "*I [as a sell-side senior analyst] routinely spend days and hours conducting industries and firms researches before writing reports; and try to promote those high-quality reports to those need them*", said by a front-line analyst when interviewed about their working patterns.

activities, opportunities and risks, etc. Despite documented effectiveness of quantitative information in both U.S. and China financial markets, (Huang et at., 2014, Yan, Wang and Kang, 2017), several problems are worth noticing: (1) quantitative measures are significantly optimistic biased and are less likely to be downward revised, due to relations with affiliating with mutual funds (Mola and Guidolin, 2009; Firth et al, 2012), career concerns (Hong and Kubik, 2003) or purely cognitive bias (Corredor et al., 2014). This is problem is profoundly augmented in China – over 92.7% of all sell-side analyst reports issuing Buy or Strong Buy; (2) discrete quantitative measures, say, recommendation level as buy or sell, contain less information than textual descriptive content (Huang et al., 2014); (3) quantitative measures cannot be directly compared without knowledge of firm-specific fundamentals. On the other side, comprehensive textual analyst reports have been proved to be informative in addition to existing quantitative measures. By evaluating and extracting tones from textual information, an investment portfolio with significant positive cumulative abnormal returns could be constructed. Hence the importance of textual content.

Important as the previously stated textual content of analyst reports, understanding information that analysts trying to convey is critical for readers. Nevertheless, analysts with distinguishing backgrounds would hardly write akin reports. Factors of analysts' characteristics, both hard characteristics and soft characteristics may contribute to this diversification. Hard qualities include educational background, industry experience, personal social network, professional certificate etc.; while soft, or intrinsic, characteristics are diversifying personality traits, such as approachability, attractiveness, youthfulness, dominance, etc. Specifically, analysts with a higher educational degree are typically taught to think critically and to think from different perspectives more; analysts with in-field industry knowledge would be able to better interpret released news and notices using past working experiences. Thus, those analysts are more than likely to uncover potential opportunities and risks. Soft characteristics are equally important: analysts with, say, high-level approachability and attractiveness may have greater probability to acquire information from social network; extrovert analysts are more likely to write out provoking, intriguing, even outrageous text, while introvert analysts may subtly conceal information into deliberate wordings. From above discussions, we make the hypothesis that:

**H1a:** Analyst characteristics, both extrinsic and intrinsic, can pose influence on qualitative textual content of analyst reports, but not on related quantitative measures.

H1a focus on effects of analysts' characteristics on general sentiment direction of analyst reports. Specific textual characteristics are also worth exploring (Huang et al., 2014). From text and paragraph perspective, we concern report title direction, report length and report conciseness; from sentence and vocabulary perspective, we concern sentence complexity, financial words and confident words usage. An intuition is that analysts with rocketing education degree and professional science experience may use plenty of text explaining the mathematic intuitions and logic behind those numbers, and analysts with multiple industry experiences may convey messages using multiple field-specific vocabularies. Both scenarios could be tremendous troubling for ordinary investors, lowering information delivering effectiveness. Here, we make following hypothesis from the discussion above:

**H1b:** Analyst characteristics can influence specific textual report characteristics.

## 3.2 Analysts' Characteristics and Prediction Accuracy

There are several reasons why analyst characteristics could explain their forecasting errors. Both extrinsic and intrinsic characteristics could, directly and indirectly, determine their prediction accuracy. First, certain extrinsic personality characteristics could directly impact analyst forecast error. Gender differences, for instance, explain itself here: multiple studies show female analysts typically are less mentored (Athey, Avery, and Zemsky, 2000) and receive fewer benefits than males from social networks (Fang and Huang, 2017). Second, hard personality characteristics could indirectly decide forecast accuracy. It is possible that years of education may lead to slow and cautious, but accurate judgments, or, equally possible, to fast but aggressive judgments. Third, individual soft characteristics may indirectly alter forecasting error, for welcomed characteristics helping them acquire more useful private information, (information discovery role), and (or) let them restate and interpret news in an easy-penetrating manner (information interpretation role). Finally, we contend that individual soft characteristics may also directly influence

forecast error: dominant and impetuous candidates, for instance, have greater chances making extreme forecasts. Based on the discussion above, we raise the hypothesis below:

**H2a:** Analyst characteristics have impacts on corresponding forecasting accuracy: welcomed (unwelcomed) characteristics could decrease (increase) forecasting error.

## 3.3 Analysts' Career Path Associated with Personality Traits

After demonstrating that certain agreeable analyst characteristics could decrease forecasting error, we next examine whether soft characteristics, i.e. personality traits, could lead to an auspicious career path. In the last quarter of each year, *New Fortune* Magazine[1] organizes the Star Analyst Election event, selecting about 10% out of all registered sell-side analysts (about 150 out of 1,500 sell-side analysts according to 2016's data) as star analysts based on votes from institutional investors including mutual fund manager, private fund managers, neutral market observers. Elected Star analysts would have rocket high annual bonus and more favorable career choices, as recent news indicating that prominent Star analyst Zepin Ren is going to be hired by a real estate giant with annual compensation more than 15 million RMB. Despite limitations of exact salary and compensation information of analysts, Star analyst is still a valid indicator for analyst career path (Bradley et al., 2017; Groysberg, Healy, and Maber, 2011). In analyst industry, being elected as Star analyst is often regarded a *must* in major trading houses – a full strategy team would all but be fired if none is elected as Star analyst in past three consecutive years[2]. Therefore, it is worth discussing whether personality traits could directly increase their social network or indirectly let them acquire more information to produce high-quality analyst reports, and further shed lights on their career outcome. Based on the logic mentioned above, we here make following hypothesis:

**H3a:** Analysts with welcomed (unwelcomed) personality traits have higher (lower) probability to have a favorable career outcome.

[1] See http://www.xcf.cn for more about *New Fortune* Magazine.
[2] As reported by a second-year sell-side analyst from a top mainland trading house.

# 4. Personality Traits Extraction and Bayesian Textual Information Classification

## 4.1 Extracting Personality Traits from Analysts' Profile Images

To extract personality traits from analysts' profile images, we collected 117,100 registered sell-side analysts profile images from 129 trading houses from *Security Association of China* (SAC)[1]. Each analyst image is named with a unique registration ID, allowing us to match their profile information. The overall personality traits extraction process can be summarized in three steps – key facial landmarks detection, facial attributes development and personality traits computation (**Figure 1**).

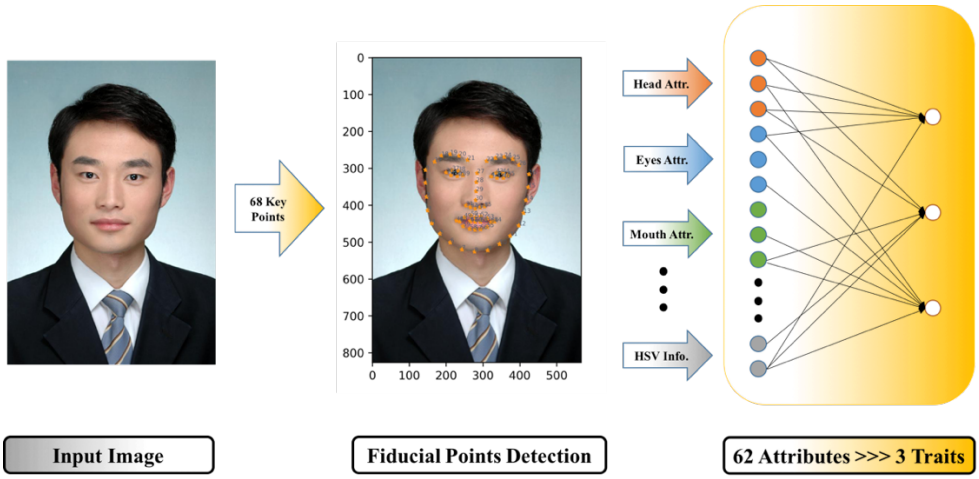



**Figure 1**

For both accuracy and efficiency concern, each image is modelled by a 68 points facial landmark detector, yielding 68 key fiducial points and a rectangular shaped area reporting the height and the width of a given image. (See **Appendix 2(a)** for the location and sequence of each point). Due to variations of image sizes and face proportion, we preprocess each image and scale face width to be 200 pixels. To analyze color information of each image, we convert each image from RGB (red, green and blue) color channels to HSV (hue, saturation and value) color channel using Python Matplotlib Library.

[1] See more at http://sac.net.cn for information about *Security Association of China*

We next derive 62 facial attributes from 68 key fiducial points. Following Vernon et al. (2014), we calculated 48 shape attributes: 6 attributes of head, 4 attributes of eyebrow, 5 attributes of eyes, 5 attributes of nose, 4 attributes of jaw, 8 attributes of mouth, 16 attributes of structural, positional and spacing features, and 14 attributes of textual and color features. Though 68 fiducial points are significantly less than 179 fiducial points manually dotted by Vernon et al. (2014), 68 points are enough to develop most facial attributes; and some quasi-measurements are developed to extract those non-directly computable attributes. Two necessary adjustments are applied to normalize measured attributes: (1) a squared root transform is applied to attributes measuring area; (2) all final attributes are normalized with range (-1,1) suggested by Vernon et al. (2014), but geometry adjustments are ignored here since images are all non-wild standard profile images and are preprocessed previously to ensure the facial width being 200 pixels. Finally, 62 (-1,1) ranged facial attributes are yielded by key points analysis mentioned above for each individual.

The *APPRO, YOAT* and *DOM*, standing for Approachability Score, Youthfulness and Attractiveness Score, and Dominance Score, respectively, are computed as linear selective combinations of those 62 attributes with fixed coefficients following Vernon et al.'s study (2014). (See **Appendix 2(b)** for selected significant explaining attributes and respecting coefficients). After computing personality scores for each person, we selectively demonstrate several analysts in top, middle, and low quintile personality scores accordingly in **Appendix Figure A.4** to vividly illustrate the idea.

## 4.2 Bayesian Textual Information Classification Approach

As a standard and effective classification approach in NLP, Naïve Bayesian classification algorithm is a statistical identification and classification approach based on prior probabilities and conditional probabilities. For sentiment uncovering task, a Naïve Bayes classifier based on pre-classified (golden) standard is trained, and then used to classify others by maximum likelihood estimation on features of language material. Formally, we have the definition of Naïve Bayes algorithm as follows:

(1) Suppose $x = \{w_1, w_2, \dots, w_m\}$ is a sentence to be classified, and each $w$ is a "feature", or word, in common language, of that sentence $x$;

(2) Suppose a category set $C = \{y_1, y_2, \dots, y_n\}$ exists, and the sentence $x$ should fall into one of those categories, i.e. $c^* \in \{y_1, y_2, \dots, y_n\}$;

(3) Calculating conditional probability of each category conditional on given sentence $x$: $P(y_1|x), P(y_2|x), \dots, P(y_n|x)$;

(4) If $c^* = \underset{c \in \{y_1, y_2, \dots, y_n\}}{argmax} P(c|x) = max\{P(y_1|x), P(y_2|x), \dots, P(y_n|x)\}$, then $x \in c^*$;

To calculate conditional probability in Step (3):

(1) Obtaining samples of sentences with known (golden) categories as training material;

(2) Calculating the conditional probability of each *feature* under each available category:

$P(w_1|y_1), P(w_2|y_1), \dots, P(w_m|y_1)$;

$P(w_1|y_2), P(w_2|y_2), \dots, P(w_m|y_2)$;

...;

$P(w_1|y_n), P(w_2|y_n), \dots, P(w_m|y_n)$;

(3) Applying Naïve Bayes algorithm with the assumption of conditional independence among all features:

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$$

(4) To maximize the above equation, we only need to maximize the numerator since the denominator $P(x)$ is a constant for all categories; With the assumption of conditional independence of all features, we have:

$$c^* = \underset{c \in \{y_1, y_2, \dots, y_n\}}{argmax} P(c) \prod_{j=1}^{m} P(w_j|c)$$

Despite the "naïve" assumption of conditional independence among features ignores structural information such as grammar and appearing sequence, NLP linguists (Lewis, 1998; Manning and Schütze, 1999) report that this assumption does not have material influences on classification accuracy when compared with classification approaches including structure information.

Domain specificity in analyst textual content is one of the strengths of Bayesian classification approach. The Naïve Bayes machine learning classification approach

learned interrelations between words and their semantic meanings in analyst specific context, thus achieves higher classification accuracy than "one-hot" traditional dictionary method. This is critical, as Pang and Lee (2008) pointed out, since certain words and phrases are highly "domain specific" and apply only to a small field when performing semantic analysis tasks.

## 4.3 Implementation of Naïve Bayes Classification Approach on Analyst Reports

**Table 1：Classification Accuracy Comparison**

| | False Pos (%) | False Neu (%) | False Neg (%) | Acc. (%) |
|---|---|---|---|---|
| **Naïve Bayes Classification Approach** | | | | |
| Manually Classified + Self-learning (In-sample Validation) | **2.0** | **3.8** | **2.1** | **92.06** |
| Manually Classified + Self-learning (Ten-folds Cross-Validation) | 5.2 | 6.0 | 4.8 | 84.26 |
| Manually Classified (In-sample Validation) | **2.0** | 3.9 | 2.8 | 91.24 |
| Manually Classified (Ten-folds Cross-Validation) | 8.3 | 11.2 | 7.0 | 73.53 |
| **Financial Dictionary Methods** | | | | |
| Word2Vec Dictionary (In-sample Validation) | 5.6 | 9.6 | 3.6 | 81.19 |
| Word2Vec Dictionary (Ten-folds Cross-Validation) | 9.9 | 20.4 | 11.6 | 58.12 |
| **General Dictionary Method** | | | | |
| General Dictionary | 7.8 | 14.5 | 10.9 | 66.81 |
| ***Boson* Sentiment Analysis API** | 39.4 | 2.3 | 3.3 | 55.10 |

This table reports classification accuracy between Naïve Bayes Approach and existing Dictionary methods.

For Naïve Bayes Classification Approach, in-sample validation accuracy reports the highest classification accuracy the model could achieve by using 10,000 manually classified sentences as both training sample and testing sample; while ten-folds cross-validation accuracy reports the robust classification accuracy by shuffling 10,000 pre-classified sentences into ten folds, with nine folds used as training sample and the remaining one used as testing sample. Accuracies in table are calculated as the average of ten repeated results from training for ten-folds cross-validation. During Training, learning rate is set to 0.001 and total number of features is set to 15,000. This setting required at least 16 GB of RAM, and takes 6 hours to complete on an Intel i5 2C4T processor.

The idea of self-learning is to increase training sample to improve model accuracy without supervision. By doubling training sample size through adding model auto-classified sentences with

confidence greater than 90% into training sample and retraining the model, both in-sample validation accuracy and ten-folds cross-validation accuracy is improved.

Dictionary Methods is the process of counting the number of positive words and negative words in each sentence by using a predefined dictionary and predicting positive if a sentence has more positive words than negative words, vice versa.

By employing Google developed Word2Vec, we randomly chose 10GB text to train the model and obtained word vectors model (Hyper Parameters: Model: Skip-Gram; dimension:100; window size: 8; iteration: 50). Cosine distance between words are obtained through this model, and similar words should be close in vector space yielding small cosine distance. Positive words dictionary and negative words dictionary are constructed by selecting top 10 most similar words for each word in the existing dictionary. Inaccuracy may due to similar vector space for both synonyms and antonyms.

Boson is a local enterprise providing real-time sentiment analysis API (http://bosonnlp.com). By calling the Python wrapped API, a tuple of positive and negative sentiment scores within range (0, 1) will be returned. The sentence sentiment score is calculated as positive score minus negative score, which lying in range (-1, 1). After multiple tests, we find that this module rarely returns sentiment scores below -0.5, thus we treat score greater than 0.5 as positive, between 0 and 0.5 as neutral and smaller than 0 as negative.

As indicated above, this probabilistic prediction model accuracy high depends on the quality of training samples since the conditional probability of each feature is obtained from training materials. By randomly selecting and carefully cross-classifying 10,000 sentences from all analyst reports either into positive, neutral or negative category, we construct our "Chinese Analyst Report Sentiment Dataset" with 2,386 sentences being positive, 6,298 sentences being neutral, and 1,316 sentences being negative. The selecting process is semi-random with category size restricted to close to Huang et al. (2014) to achieve better classification accuracy. Scikit-Learn[1] is employed here to learn pre-labelled sentences. The Naïve Bayes classifier accuracy is compared with traditional dictionary classification accuracies in **Table 1**.

The accuracy comparison report above indicates that the Naïve Bayes classifier achieves the highest classification accuracy among other traditional methods with in-sample validation accuracy of 92.06% and ten-folds cross-validation accuracy of 84.26% by augmenting training samples through self-learning. Confusion matrix indicates that major false results locate in "False Neutral" area while minor in "False Positive" and "False Negative" areas. Therefore, we contend that those inaccurate classification result will not jeopardize classifier's overall performance, and our Naïve Bayes classification approach is reasonably accurate and robust.

[1] Scikit-learn, a machine learning specific Python module open sources with BSD license, see more at http://www.scikit-learn.org; training feature size is 15,000, training learning rate is 0.001

## 4.4 Measurement of Textual Opinion

By inputting each sentence from a given analyst report to the trained classifier, we obtain the number of positive sentences (*N_POS*), the number of neutral sentences (*N_NEU*) and the number of negative sentences (*N_NEG*). The summation of above three variables is the total length[1] (*LENGTH = N_POS+ N_NEU+N_NEG*) of an analyst report. The tone, or *OPINION* level, of a given analyst report is calculated as the difference between the percentage of positive sentences (*POS_PCT*) and the percentage of negative sentences (*NEG_PCT*). Formally, we have:

$$OPN = \frac{N\_POS\ - N\_NEG}{LENGTH} = POS\_PCT - NEG\_PCT \quad (1)$$

Noticing that more positive information content than negative information is witnessed in analyst reports and the mean of *OPN* is larger than 0, which means that the total sentiment level *OPN* would decrease with increased number of neutral sentences holding the number of positive and negative sentences constant. In other words, ceteris paribus, neutral sentences could dilute the overall sentiment level, *OPN*.

# 5. Sample Selection and Variable Definition

## 5.1 Selection of Analyst Reports and Profile Images

We employ Python Scrapy module to massively collect massive analyst reports from main financial information providers *Tencent Finance* and *East Money*; HTML (Hypertext Markup Language) formatted reports allow us to extract report dates, report titles, analyst names, affiliated trading houses, and full contexts of reports through analyzing tagged information. From a starting sample of 1,068,679 analyst reports from 2006 to 2016, we delete market research reports, industry research reports and others with multiple referred firm, remaining 309,056 reports. We then match our reports with

[1] For each analyst report, textual contents are cleaned to remove meaningless information such as disclaimers, brokerage description, etc. Details on cleaning analyst report content can be found in Appendix 1

CSMAR Analyst database to obtain recommendation level (*REC*), recommendation revision (*REC_CNG*), Forecast EPS (*FEPS*) as well as financial market cumulative abnormal returns (*CAR*), giving our 165,820 samples remaining. To further combine analysts' extrinsic characteristics, we merge SAC analyst information with our analyst reports to get 110,830 samples. We finally merge analyst intrinsic characteristics with corresponding reports by analyst name and affiliated trading house. After filtering out images with zero or multiple detected faces, and images with extremely low resolution – image size less than 10KB, we manage to get a sub-sample of 12,374 analyst reports with detailed analyst profile information and market information.

## 5.2 Description of Analyst Reports and Profile Images

**Table 2** provides descriptive statistics of main variables with **Panel A** describing analyst report level information, **Panel B** and **Panel C** describing analysts' extrinsic and intrinsic characteristics variables, respectively. Generally, we observe the average length of each analyst report is 20 sentences, with a perceptible decreasing trend exhibited in **Figure 2**. This is consistent with research conducted by *New Fortune* (2017) that mobile reading encourages short, concise and straight-forward message delivery way, rather than conventional long, deep thorough analysis. Overall, analyst reports convey much more positive information (72.54%) than negative information (10.51%), with total *OPN* level being 62.03%, consistent with analysts' optimism. Yet this number is significantly higher than U.S. data with the mean around 18% (Huang et al., 2014). Nevertheless, it is not due to our Naïve Bayes classification approach, for 63% of training data being neutral and 28% of training data being positive – the classifier could have achieved higher accuracy by leaning towards neutral category. Further, analyst quantitative measures along with reports support above mentioned observation: 92% of analyst report issue Strong Buy or Buy recommendation, significantly higher than U.S. analysts' 55%; while only 0.25% China analysts issued Sell or Strong Sell during the past ten years, yet 6% is observed in U.S. samples (Huang et al., 2014). Similar patterns are found for other quantitative level and revision measures. (See **Appendix 1(a)** for detailed comparison over China and U.S. analyst reports' recommendation and revision).

**Table 2: Summary Statistics of Main Variables**

**Panel A: Analyst Reports Textual Opinion**

| | Obs. | Mean | Std. Dev. | Min | Median | Max |
|---|---|---|---|---|---|---|
| *OPN* | 165,820 | 0.6203 | 0.2147 | -0.9000 | 0.6500 | 0.9737 |
| *LENGTH* | 165,820 | 19.7578 | 11.3492 | 3.0000 | 17.0000 | 63.0000 |
| *POS_PCT* | 165,820 | 0.7254 | 0.1433 | 0.0000 | 0.7500 | 0.9737 |
| *NEU_PCT* | 165,820 | 0.1696 | 0.1225 | 0.0189 | 0.1304 | 1.0000 |
| *NEG_PCT* | 165,820 | 0.1051 | 0.1000 | 0.0000 | 0.0833 | 0.9000 |
| *REC_LEVEL* | 155,617 | 4.3512 | 0.6265 | 1.0000 | 4.0000 | 5.0000 |
| *REC_CNG* | 134,229 | 2.0181 | 0.2431 | 1.0000 | 2.0000 | 3.0000 |
| *FEPS* | 164,689 | 0.8648 | 0.9233 | -2.8900 | 0.6700 | 19.7400 |
| *FEPS_CNG* | 165,800 | 0.0035 | 0.2626 | -16.3000 | 0.0000 | 18.3700 |
| *FTP* | 34,510 | 30.3760 | 28.8012 | 23.4000 | 0.0000 | 430.0000 |
| *FTP_CNG* | 21,403 | 0.3704 | 15.4101 | -220.0000 | 0.0000 | 220.0000 |
| *PRIOR_CAR* | 252,844 | 0.0160 | 0.1058 | -0.5075 | 0.0077 | 16.7705 |
| *CAR0_1* | 136,780 | 0.0075 | 0.1417 | -0.1248 | 0.0000 | 0.1827 |
| *CAR0_7* | 136,780 | 0.0087 | 0.1377 | -0.1810 | 0.0000 | 0.2814 |
| *CAR0_14* | 136,780 | 0.0105 | 0.1586 | -0.2379 | -0.0009 | 0.3601 |
| *CAR0_30* | 136,780 | 0.0170 | 0.1895 | -0.3004 | -0.0019 | 0.5295 |
| *BM* | 162,868 | 1.2223 | 2.3288 | 0.0004 | 0.5464 | 143.8035 |
| *TRSIZE* | 125,908 | 23.9071 | 1.2641 | 13.3136 | 24.1507 | 27.3824 |
| *SIZE* | 161,459 | 19.7970 | 1.7068 | 10.3382 | 19.5605 | 26.3549 |
| *ROA* | 165,415 | 0.0651 | 0.0541 | -3.2679 | 0.0575 | 0.6362 |

*OPN* = analyst reports textual opinion, calculated as *POS_PCT – NEG_PCT*;
*LENGTH* = the number of sentences of an analyst report, winsorized at 1% level;
*POS_PCT* = the percentage of positive sentences in an analyst report; each sentence is classified by pre-trained Naïve Bayes classifier model;
*NEU_PCT* = the percentage of neutral sentences in an analyst report; each sentence is classified by pre-trained Naïve Bayes classifier model;
*NEG_PCT* = the percentage of negative sentences in an analyst report; each sentence is classified by pre-trained Naïve Bayes classifier model;
*REC_LEVEL* = analyst recommendation level, 1: Sell and Strong Sell; 2: Hold; 3: Buy and Strong Buy; data is available on CSMAR database;
*REC_CNG* = analyst recommendation revision, 1: Downgrade; 2: Reiteration; 3: Upgrade;
*FEPS* = analyst forecasted EPS; data is available on CSMAR database;
*FEPS_CNG* = = analyst forecasted EPS revision, calculated as the current *FEPS* minus previous *FEPS* for the same covered firm by the same analyst;
*FTP* = analyst forecasted Target Price; data is available on CSMAR database;
*FTP_CNG*= = analyst forecasted *FTP* revision, calculated as the current *FTP* minus previous *FTP* for the same firm by the same analyst;
*PRIOR_CAR* = ten-day cumulative abnormal return prior to issue date of analyst report; abnormal return is calculated as difference between stock return and Fama-French portfolio market average return;
*CAR0_1* = two-day cumulative abnormal return from to issue date of analyst report; abnormal return is calculated as difference between stock return and Fama-French portfolio market average return;
*CAR0_7* = one-week cumulative abnormal return from to issue date of analyst report; abnormal return is calculated as difference between stock return and Fama-French portfolio market average return;

*CAR0_14* = two-week cumulative abnormal return from to issue date of analyst report; abnormal return is calculated as difference between stock return and Fama-French portfolio market average return;
*CAR0_30* = one-month cumulative abnormal return from to issue date of analyst report; abnormal return is calculated as difference stock realized return and Fama-French portfolio market average return;
*BM* = Book-to-Market value of the covered firm;
*TRSIZE* = the logarithm of net capital of brokerage house;
*SIZE* = the logarithm of last year's total assets of the covered firm;
*ROA* = the last year's return on asset ratio of the covered firm;

**Panel B: Analyst Characteristics**

| | Obs. | Mean | Std. Dev. | Min | Median | Max |
|---|---|---|---|---|---|---|
| *SITE* | 165,820 | 0.0350 | 0.1839 | 0.0000 | 0.0000 | 1.0000 |
| *FEMALE* | 118,956 | 0.3059 | 0.4608 | 0.0000 | 0.0000 | 1.0000 |
| *GROUP* | 165,800 | 1.3603 | 0.5758 | 1.0000 | 1.0000 | 3.0000 |
| *MIX* | 118,956 | 0.0732 | 0.2605 | 0.0000 | 0.0000 | 1.0000 |
| *STAR* | 281,884 | 0.1378 | 0.3447 | 0.0000 | 0.0000 | 1.0000 |
| *EDU* | 113,397 | 2.0318 | 0.4591 | 0.0000 | 2.0000 | 3.0000 |

*SITE* = 1 if analysts had site-visited to the covered firm in analyst report within 20 business days; 0 otherwise;
*FEMALE* = 1 if analyst is female or one of analysts is female; 0 otherwise;
*GROUP* = 1 if analyst work along; 2 if two analysts work together; 3 if three or more analysts work together;
*MIX* = 1 if analysts work as a team with both male and female; 0 otherwise;
*STAR* = 1 if analysts is star analyst selected by *New Fortune*; 0 otherwise;
*EDU* = 1 if analysts obtained undergraduates degree or lower; 2 if obtained graduated degree; 3 if obtained doctor degree or higher;

**Panel C: Analyst Personality Traits:**

| | Obs. | Mean | Std. Dev. | Min | Median | Max |
|---|---|---|---|---|---|---|
| *APPRO* | 29,061 | 2.7221 | 13.2215 | -31.8423 | 4.1503 | 37.1052 |
| *YOAT* | 24,198 | 3.2742 | 0.5957 | 2.2246 | 3.1418 | 4.8523 |
| *DOM* | 29,472 | -1.0697 | 0.3433 | -1.9459 | -1.0501 | -0.3915 |

*APPRO* = standardized Approachability score; calculated as linear combination of approachability personality factors by modelling 68 fiducial points;
*YOAT* = standardized Youthfulness and Attractiveness score; calculated as linear combination of youthfulness and attractiveness personality factors by modelling 68 fiducial points;
*DOME* = standardized Dominance score; calculated as linear combination of dominance personality factors by modelling 68 fiducial points;

Summary of extrinsic characteristics variables reveal some interesting facts about China's analysts' universe. About 3.5% of all analyst reports are rolled out with site-visit, recorded visit up to 20 business days prior to report release date. Female analysts account for 30% of the total. Team working is encouraged with 20.61% reports are finished by two or more analysts; and about 7% of are jointly done by at least one male and one

female analyst. Star analysts wrote about 13% of all analyst reports and this field like individuals with high level of education – 80.03% obtained master's degrees and 11.78% obtained doctor's degrees or post-docs.

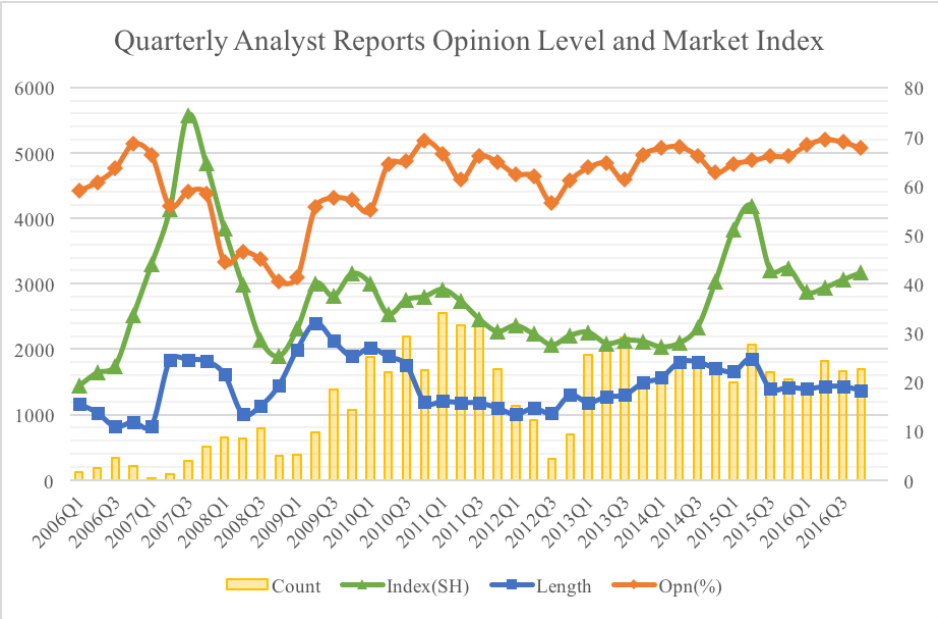



**Fig. 2**

**Figure 2** plots the trend of analyst reports *OPN* level, reports' length and the number of reports released quarterly with Shanghai Stock Composite Index. The average *LENGTH* and *OPN* of analyst report are positively correlated with composite index, with an observed surge during 2007 and 2015's bull market and significant drop during followed jumps. Analysts choose to release more reports when the stock index is booming but choose to remain silent when the market is stagnant or slipping; this is supported by the empirical report of Chen et al. (2011) that poor relative performances are noticed when executives' stop giving earning guidance. The length of reports is *not* as seasonally trended as being observed in listed company's notices (Yan, Wang and Kang, 2017) and a typical report main content consists about 20 sentences to summarize previous operating, investing and financing events, and to analyze current opportunities and risks. Short and concise general reports, usually kept within 2-3 pages, are more discerned in recent year majorly due to an increased popularity of mobile reading habits and increased competition in sell-side analyst field. (*New Fortune* Magazine, 2017). Fluctuation in textual analyst reports are diminishing with time, partly due to the increased number of homogeneous

reports. Thus, it is worth discussing whether report contents could still provide additional information content and whether analyst characteristics could affect content and related prediction accuracy.

# 6. Empirical Results

## 6.1 Analyst Characteristics and Report Textual Content

To examine how much soft information and hard information could be triggered by per unit change of analyst characteristics beyond firm and market level controls, we construct the following multivariate regression:

$$OPN = \alpha_0 + \alpha_1 SITE + \beta_1 APPRO + \beta_2 YOAT + \beta_3 DOM + \gamma_1 FEMALE + \gamma_2 EDU + \gamma_3 STAR + \gamma_4 TRSIZE + \sum_j \delta_j Controls_j + \varepsilon \quad (2)$$

We include both analyst soft characteristics and hard characteristics in our estimation Equation (2) since existing literatures document that certain analyst level characteristics have impacts on investors at financial market level (Huang et al., 2014, Bradley et al., 2017). In Equation (2), *OPN* is the general textual sentiment level generated by Naïve Bayes algorithm, with detailed descriptions in previous section. Site-visit (*SITE*) indicates whether analysts conduct site-visits to the related firm within 20 business days. *APPRO*, *YOAT*, and *DOM*, abbreviations for Approachability Score, Youthful and Attractiveness Score, and Dominance Score, respectively, evaluate analysts level intrinsic characteristics, which are normalized within range (-1,1) with detailed construction process in previous section; while *FEMALE*, *EDU*, *STAR*, and *TRSIZE* measure analyst gender, education level, whether being current year Star analyst, and affiliated trading house size. *FEMALE* is a dummy variable, being 1 if analyst, or at least one analyst from a research team, is female; *EDU* level reveals group average education level. For each analyst, education level (*EDU*) takes 1 if being undergraduate or below; takes 2 if being master; and takes 3 if being doctor or above. To further investigate the impact of analyst characteristics on different types of information, we also replace dependent variable from soft information to hard information, recommendation level

change (*REC_CNG*), where equals 3 if upgraded, 2 if reiterated and 1 if downgraded, and forecasted EPS change (*FEPS_CNG*), calculated as current forecasted EPS ($FEPS_t$) minus previous forecasted EPS ($FEPS_{t-1}$) issued by the same analyst for the same firm, adjusted by stock price. Above mentioned analyst information is publicly disclosed in SAC's (Security Association of China) website. *STAR* is a dummy representing whether being elected as current year's Star analyst by *New Fortune* Magazine. Affiliated trading house size, *TRSIZE*, reveals analyst market dominance, available industry resources, potential social network and career outcome, and is measured by the natural logarithm of last year's net profits. To match analyst characteristics measures with textual content, we rigorously require a full match of analyst name, affiliated trading house, and report releasing year[1].

Several control variables are included in regression Equation (2). Considering analysts report could be an aftermath of recent announced news and notices, we include ten-day cumulative abnormal return prior to the report releasing date (*PRIOR_CAR*). For documented information content of analyst report in Huang et al. (2014), *PRIOR_CAR* controls for short-term momentum or reversal between stock price and analyst textual report content. To control for analyst's textual deviations due to market level and firm level variations, we add firm return-on-asset ratio (*ROA*), the natural logarithm of covered firm's total assets (*SIZE*), firm Book-to-Market ratio (*BM*) and industry and year fixed effect into our estimation Equation (2). Standard errors are estimated by two-way clustering at analyst and firm levels, for each analyst may follow multiple firms and each firm may be followed by multiple analysts.

**Table 3** reports estimation result of Equation (2). Column (1) reports estimation of analyst characteristics on soft information, *OPN*; while column (2) and column (3) report results on hard information – analyst recommendation level change (*REC_CNG*) and

[1] A partial match of analyst name, and social year, excluding affiliated trading house name, is also performed as a robustness check, which increases samples by about 20%. The increased samples are from two channels: (1) analysts with same names but in different brokerage houses are cross-matched in partial matching criteria. This could either be caused by two analysts with same name working for two different brokerages, or be caused by a single analyst changing occupation during career. These two plausible and normal phenomena are not easily distinguished without SAC unique registration number; (2) brokerages' abbreviation names may change, though the official recorded full names are unique – making it challenging to match. For example, ShenWanYanJiu shall be treated as one trading house as ShenWanZhengQuan since it is merely a sub-research team of ShenWan Group. Albeit inaccurate cross-match could occur during the partial match process, estimation results, both significance level and economical meaning, are similar.

analyst forecasted EPS change (*FEPS_CNG*). Comparing column (1) with (2) and (3), we find consistent estimation results with hypothesis H1a, which argues that both hard and soft analyst characteristics could impact qualitative textual content, but not on quantitative hard measures: a number of analyst characteristics informatively shad lights on soft *OPN*, but none is significant on hard information. Firstly, analysts' site-visit is not only statistically significant (P<0.01) but also economically informative – post site-visit Opinion level are observed to be augmented by 7.88%, which means more number of positive arguments could be found after site-visit; but, interestingly enough, analysts choose to not adjust recommendation level (*REC*) or forecasted EPS (*FEPS*). Secondly, Soft personality characteristics are significantly correlated with report Opinion level. One standard deviation increase of Youthfulness and Attractiveness and Dominance score yields an upper shifted general textual Opinion level by 0.78% and 0.44%, respectively. Approachability is statistically negatively correlated with Opinion level, yet it's economical meaning is trivial. Thirdly, among those listed extrinsic characteristics, education and trading house size have correlations with report Opinion – one higher degree level decease Opinion level by 1.12% and one standard deviation increase of log net profits of trading house increase Opinion level by 0.32%. Analyst team characteristics and whether being elected as Star analyst are not informative in neither case[1].

[1] In fact, we may safely conclude that team-level characteristics are not informative. we examined analyst team compositions: *FEMALE*, whether have at least one female analyst; *GROUP*, whether have more than one analyst; and *MIX*, whether have at least one male and one female analyst. Those team-level characteristics variables are not significant by either individually or jointly added into Equation (2), and Adjusted R-square only increases by less than 0.1% from initial 12.67%.

**Table 3: Analysts Characteristics and Textual Opinion**

**Panel A: Analysts Characteristics on Soft Information and Hard Information**

| | (1) OPN Coefficients (t-stats) | (2) REC_CNG Coefficients (t-stats) | (3) FEPS_CNG Coefficients (t-stats) |
|---|---|---|---|
| *SITE* | 0.0788*** | 0.0045 | 0.0108 |
| | (9.49) | (0.48) | (1.24) |
| *APPRO* | -0.0009*** | 0.0002 | -0.0002 |
| | (-3.84) | (1.10) | (-0.94) |
| *YO_AT* | 0.0136*** | 0.0017 | 0.0080 |
| | (2.65) | (0.36) | (1.35) |
| *DOM* | 0.0222*** | 0.0048 | 0.0079 |
| | (2.64) | (0.59) | (1.00) |
| *FEMALE* | -0.0012 | -0.0034 | -0.0048 |
| | (-0.18) | (-0.55) | (-0.74) |
| *EDU* | -0.0123** | -0.0029 | -0.0056 |
| | (-2.45) | (-0.53) | (-0.92) |
| *STAR* | 0.0036 | -0.0036 | -0.0046 |
| | (0.62) | (-0.69) | (-0.81) |
| *TRSIZE* | 0.0025* | 0.0017 | -0.0008 |
| | (1.92) | (0.96) | (-0.61) |
| *PRIOR_CAR* | 0.0421** | 0.0369 | 0.1032*** |
| | (1.97) | (1.33) | (3.37) |
| *ROA* | 0.2181*** | 0.1470** | 0.4083*** |
| | (3.16) | (2.51) | (3.38) |
| *SIZE* | 0.0015 | -0.0013 | 0.0034 |
| | (0.67) | (-0.79) | (1.50) |
| *BM* | -0.0011 | 0.0023 | 0.0040** |
| | (-0.61) | (1.41) | (2.03) |
| *Cons.* | 0.5314*** | 2.6451*** | -0.0980* |
| | (8.59) | (9.22) | (-1.65) |
| *Adj. $R^2$* | 12.72% | 1.30% | 1.58% |
| *Obs.* | 12,203 | 10.545 | 12,203 |

***, ** and * indicate significance level at 1%, 5% and 10%, respectively;
All estimations have controlled Year and Industry fixed effect;
Standard Errors are estimated by two-way cluster at firm and analyst level;

**Panel B: Analysts Characteristics and Personality Traits on Textual Opinion**

| | (1) OPN Coefficients (t-stats) | (2) OPN Coefficients (t-stats) | (3) OPN Coefficients (t-stats) | (4) OPN Coefficients (t-stats) | (5) OPN Coefficients (t-stats) |
|---|---|---|---|---|---|
| *SITE* | 0.0788*** | 0.0788*** | 0.0788*** | 0.0788*** | 0.0787*** |
| | (9.83) | (9.49) | (9.82) | (9.49) | (9.47) |
| *APPRO* | -0.0010*** | -0.0009*** | -0.0010*** | -0.0009*** | -0.0009*** |
| | (-4.44) | (-3.84) | (-4.42) | (-3.98) | (-3.86) |
| *YOAT* | 0.0130** | 0.0136*** | 0.0132*** | 0.0138*** | 0.0137*** |
| | (2.52) | (2.65) | (2.58) | (2.71) | (2.68) |
| *DOM* | 0.0215*** | 0.0222*** | 0.0216*** | 0.0218*** | 0.0225*** |
| | (2.65) | (2.64) | (2.66) | (2.66) | (2.69) |
| *FEMALE* | | -0.0012 | | | 0.0023 |
| | | (-0.18) | | | (0.34) |
| *GROUP* | | | -0.0032 | | -0.0014 |
| | | | (-0.59) | | (-0.22) |
| *MIX* | | | | -0.0098 | -0.0098 |
| | | | | (-1.09) | (-0.85) |
| *EDU* | -0.0127** | -0.0123** | -0.0122** | -0.0114** | -0.0115** |
| | (-2.57) | (-2.45) | (-2.43) | (-2.29) | (-2.27) |
| *STAR* | 0.0040 | 0.0036 | 0.0042 | 0.0041 | 0.0045 |
| | (0.70) | (0.62) | (0.75) | (0.70) | (0.77) |
| *TRSIZE* | 0.0025* | 0.0025* | 0.0027** | 0.0027** | 0.0027** |
| | (1.92) | (1.92) | (2.00) | (2.04) | (2.04) |
| *PRIOR_CAR* | 0.0479** | 0.0421** | 0.0477** | 0.0418* | 0.0416* |
| | (2.25) | (1.97) | (2.24) | (1.96) | (1.95) |
| *ROA* | 0.2073*** | 0.2181*** | 0.2069*** | 0.2168*** | 0.2163*** |
| | (3.02) | (3.16) | (3.02) | (3.15) | (3.15) |
| *SIZE* | 0.0018 | 0.0015 | 0.0018 | 0.0016 | 0.0016 |
| | (0.82) | (0.67) | (0.82) | (0.70) | (0.71) |
| *BM* | -0.0013 | -0.0011 | -0.0013 | -0.0011 | -0.0012 |
| | (-0.71) | (-0.61) | (-0.72) | (-0.62) | (-0.64) |
| *Cons.* | 0.5262*** | 0.5314*** | 0.5239*** | 0.5213*** | 0.5221*** |
| | (8.56) | (8.59) | (8.51) | (8.37) | (8.39) |
| *Adj. $R^2$* | 12.62% | 12.72% | 12.62% | 12.73% | 12.73% |
| *Obs.* | 12,374 | 12,203 | 12,374 | 12,203 | 12,203 |

***, ** and * indicate significance level at 1%, 5% and 10%, respectively;
All estimations have controlled Year and Industry fixed effect;
Standard Errors are estimated by two-way cluster at firm and analyst level.

Two possible explanations are proposed here for observations above: (1) more than 92.7% of China analyst reports are already in Buy or Strong Buy position, leaving analysts with little space for further upgrading; (2) hard information, such as forecasted EPS (*FPES*) and recommendation (*REC*), could easily be back-examined: unrealized EPS upgrade could be viewed as hard evidence as incompetence and insufficient of analysis

skills, shadowing analyst career path. Thus, analysts are more than likely to adjust soft, implicit, information than hard, explicit, information[1].

To Test H1b, we decompose general Opinion and focus on six specific textual characteristics. *TITLE* is opinion level of report title, where equals 1 if being positive, 0 if being neutral and -1 if being negative. *LENGTH* measure the number of sentences of a given document, *COMPLEX* measures the average of the number of Chinese characters per sentence; both variables are preprocessed with winsorization at 1% level[2]. *FIN* evaluate the financial related information discussed level and is measured as the number of sentences containing RMB currency symbol or percentage symbol ("￥" or "%"). *CONFI* measures the general confidence level of an analyst report, counting the number of sentences containing words exhibiting strong confidence emotion. *CONCISE* measure the conciseness of a given document and is derived following Huang et al. (2014) by calculating -1 times the residual from regressing report length (*LENGTH*) on firm size (*SIZE*), the book-to-market ratio (*BM*), and the recent returns (*PRIOR_CAR*). Summary of above-mentioned variables are made available in **Table 4 Panel A**. We accordingly design our multivariable regression model as Equation (3)

$$\begin{aligned} &Textual\ Characteristics \\ &= \alpha_0 + \alpha_1 SITE + \beta_1 APPRO + \beta_2 YOAT + \beta_3 DOM + \gamma_1 FEMALE \\ &+ \gamma_2 EDU + \gamma_3 STAR + \gamma_4 TRSIZE + \sum_j \delta_j Controls_j + \varepsilon \end{aligned} \quad (3)$$

[1] This argument is proved by a junior stock analyst, with his quote, "[as stock analysts], we would use predefined templates when writing reports, and seldom change recommendations, especially to downgrades, since it would ruin our relationship with firm's management team". He also pointed out that it is text, rather than numbers (forecasted EPS, target price), matters.

[2] We winsorize both *LEGNTH* and *COMPLEX* at 1% level to avoid extreme numbers. We winsorize *LEGNTH* and *COMPLEX* because numbers of "fake sentences" may be generated by tokenizer if a document has tables containing Chinese characters. Another reason for us to winsorize *COMPLEX* is that sentence ending token may be truncated due to conflicts during data-crawling, persistence storing, or sophisticated NLP analysis.

**Table 4: Analyst Characteristics and Textual Content Characteristics**

**Panel A: Summary Statistics**

| | Obs. | Mean | Std. Dev. | Min | Median | Max |
|---|---|---|---|---|---|---|
| *TITLE* | 165,810 | 0.5362 | 0.7667 | -1.0000 | 1.0000 | 1.0000 |
| *LENGTH* | 165,820 | 19.7578 | 11.3492 | 3.0000 | 17.0000 | 63.0000 |
| *FIN* | 165,820 | 2.9077 | 2.1292 | 0.2414 | 2.4707 | 22.3565 |
| *COMPLEX* | 165,820 | 63.0143 | 22.8885 | 31.6400 | 57.8462 | 167.0000 |
| *CONFI* | 165,820 | 2.3857 | 1.6524 | 0.0000 | 2.1007 | 23.6662 |
| *CONCISE* | 136,780 | -8.3498 | 8.5547 | -282.2076 | -6.2242 | -0.0001 |

*TITLE* = the opinion level of analyst report; 1 = positive, 0 = neutral, -1 = negative;
*LENGTH* = the number of sentences of an analyst report, winsorized at 1% level;
*FIN* = the number of financial words divided by natural logarithm of *LENGTH* of an analyst report;
*COMPLEX* = the average of the number of words per sentence of an analyst report;
*CONFI* = the number of confident words divided by natural logarithm of *LENGTH* of an analyst report;
*CONCISE* = conciseness level of analyst report, calculated as the *negative* absolute value of residual estimated by regressing *LENGTH* on *PRIOR_CAR, BM,* and *SIZE* following Huang et al (2014);

**Panel B: Estimation Results**

| | TITLE | LENGTH | CONCISE | FIN | COMPLEX | CONFI |
|---|---|---|---|---|---|---|
| | (1) Coefficients (t-stats) | (2) Coefficients (t-stats) | (3) Coefficients (t-stats) | (4) Coefficients (t-stats) | (5) Coefficients (t-stats) | (6) Coefficients (t-stats) |
| *SITE* | 0.1318*** | 1.0541** | -0.1830 | -0.5869*** | -0.2214 | 0.1000 |
| | (4.63) | (2.17) | (-0.53) | (-6.63) | (-0.14) | (1.06) |
| *APPRO* | -0.0002 | 0.0356*** | -0.0206** | 0.0040* | 0.0085 | 0.0003 |
| | (-0.25) | (2.88) | (-2.33) | (1.71) | (0.38) | (0.15) |
| *YO_AT* | 0.0069 | -0.0162 | -0.3110* | -0.2618*** | -0.8583 | 0.0471 |
| | (0.34) | (-0.06) | (-1.68) | (-4.86) | (-1.59) | (1.10) |
| *DOM* | 0.0918*** | -0.6296 | 0.7273** | -0.1348 | -0.2355 | 0.1600** |
| | (2.87) | (-1.46) | (2.52) | (-1.22) | (-0.21) | (2.35) |
| *STAR* | 0.0251 | -0.9571*** | -0.0274 | 0.0368 | 5.2141*** | 0.1250** |
| | (1.20) | (-3.18) | (-0.14) | (0.58) | (5.78) | (2.31) |
| *FEMALE* | 0.0427* | 0.5318 | 0.2328 | 0.3615*** | -3.8749*** | -0.1402** |
| | (1.68) | (1.52) | (0.92) | (4.70) | (-3.87) | (-2.57) |
| *EDU* | 0.0255 | 1.2470*** | -0.4878** | 0.2612*** | -1.0548 | -0.0226 |
| | (1.39) | (4.36) | (-2.48) | (4.00) | (-1.40) | (-0.51) |
| *TRSIZE* | 0.0297*** | -0.0085 | -0.2717*** | -0.0740*** | -0.1079 | -0.0414*** |
| | (4.98) | (-0.13) | (-6.12) | (-4.17) | (-0.82) | (-4.59) |
| *PRIOR_CAR* | 0.1578** | 1.6262 | -1.9520** | -0.6861*** | 0.5790 | 0.2137 |
| | (1.96) | (1.46) | (-2.38) | (-3.35) | (0.28) | (1.13) |
| *ROA* | 0.1164 | -0.0303 | 1.0087 | 1.2238 | -0.5632 | 1.1989** |
| | (0.46) | (-0.01) | (0.42) | (1.57) | (-0.06) | (2.15) |
| *SIZE* | -0.0094 | 0.0544 | -0.1532* | 0.0974*** | 0.4998 | 0.0033 |
| | (-1.09) | (0.48) | (-1.95) | (3.48) | (1.27) | (0.17) |
| *BM* | -0.0114 | -0.1445* | -0.0679 | 0.0159 | 0.2718* | -0.0166 |
| | (-1.64) | (-1.88) | (-1.14) | (0.66) | (1.94) | (-1.23) |
| *Cons.* | 0.2069 | 8.8363*** | 2.6565 | 1.8716*** | 26.5463*** | 3.2900*** |
| | (0.86) | (2.93) | (1.29) | (2.60) | (3.30) | (6.69) |
| *Adj.* $R^2$ | 3.45% | 15.18% | 9.15% | 14.58% | 10.32% | 19.03% |
| *Obs.* | 12,203 | 12,203 | 12,203 | 12,203 | 12,203 | 12,203 |

***, ** and * indicate significance level at 1%, 5% and 10%, respectively;
All estimations have controlled Year and Industry fixed effect;
Standard Errors are estimated by two-way cluster at firm and analyst level.

Six columns of **Table 4 Panel B** provide effects of analyst characteristics to corresponding textual characteristics. After conducting site-visitation, longer, more optimism biased textual reports are found. This could happen because easy-understanding non-financial contents, less inaccessible numbers and models, are discussed. Impenetrable information is interpreted as negative news since it is challenging for ordinary investors to judge ambiguous information quality. Junior analysts tend to carry out reports more from operating perspective than financial analysis; dominant analysts

tend to write assertive reports with exceedingly upper biased headlines; conservation is again observed in female analysts group with significantly less usage of confidence words.

In all, these additional detailed textual characteristics test provides us with channels to justify the impact of analyst intrinsic and extrinsic characteristics on general report Opinion level.

## 6.2 Analyst Characteristics and Prediction Accuracy

After documenting the relationship between analyst characteristics and textual content, a natural research question is whether those characteristics have market implications. We employ forecast error (*FE*) and absolute forecast errors (*AFE*) to evaluate relative prediction accuracy by calculating forecasted EPS minus realized EPS, with standardization by market price. **Panel A** in **Table 5** illustrates descriptive statistics of these dependent variable mentioned above. The mean of *FE* and *AFE* are 2.25% and 2.30% per unit currency per share, all greater than 0, which means analyst forecast EPS is statistically higher than realized EPS, accords with analyst optimistic bias in previous studies (Mola and Guidolin, 2009; Firth et al, 2012; Hong and Kubik, 2003; Corredor et al., 2014). We then design the following multivariate regression:

$$(Absolute)\ Forecast\ Error = \alpha_0 + \alpha_1 OPN + \alpha_1 SITE + \alpha_1 OPN{\times}SITE + \gamma_1 FEMALE + \gamma_2 EDU + \gamma_3 STAR + \gamma_4 TRSIZE + \sum_j \delta_j Controls_j + \varepsilon \quad (4)$$

In Equation (4), we include general report Opinion (*OPN*) and the interaction term of report Opinion (*OPN*) and site visitation dummy (*SITE*), given proved information content of textual reports in financial market to cumulated abnormal return (*CAR*) (Huang et al., 2014; Yan, Wang and Kang, 2017). Analyst soft facial traits are removed in this model since: (1) analyst profile images and intrinsic analyst characteristics are not directly observable in financial market since analyst images are not a required component and seldom appear in analyst reports. Thus, we contend that analyst personality traits are concealed in their text report and shall be measured be Opinion (*OPN*); (2) time-invariant profile image from SAC website restricts us from measuring individual time variances, and, accordingly, projecting a constant profile image to their performance could be inaccurate and bias. We also include control variables as in Equation (2).

**TABLE 5: Effects of Analyst Characteristics on Prediction Accuracy**

**Panel A: Summary Statistics**

| | Obs. | Mean | Std. Dev. | Min | Median | Max |
|---|---|---|---|---|---|---|
| *FE* | 114,470 | 0.0225 | 0.0204 | -0.0120 | 0.0173 | 0.1098 |
| *AFE* | 114,470 | 0.0230 | 0.0200 | 0.0000 | 0.0173 | 0.1098 |

*FE*= the standardized unexpected earning, calculated as the forecasted EPS minus realized EPS standardized by stock price with winsorized at 1% level;

*AFE* = the absolute value of standardized unexpected earning, calculated as the EPS minus forecasted EPS standardized by stock price; winsorized at 1% level;

**Panel B: Estimation Results**

| | (1) FE Coefficients (t-stats) | (2) FE Coefficients (t-stats) | (3) FE Coefficients (t-stats) | (4) AFE Coefficients (t-stats) | (5) AFE Coefficients (t-stats) | (6) AFE Coefficients (t-stats) |
|---|---|---|---|---|---|---|
| *OPN* | -0.0010* | -0.0010* | -0.0009* | -0.0013*** | -0.0013*** | -0.0013*** |
| | (-1.89) | (-1.87) | (-1.79) | (-3.15) | (-3.12) | (-3.04) |
| *SITE* | 0.0006 | 0.0005 | 0.0006 | 0.0003 | 0.0002 | 0.0003 |
| | (0.60) | (0.47) | (0.58) | (0.50) | (0.33) | (0.48) |
| *OPN * SITE* | 0.0021* | 0.0022* | 0.0021* | 0.0024*** | 0.0025*** | 0.0024*** |
| | (1.69) | (1.80) | (1.68) | (2.65) | (2.78) | (2.64) |
| *FEMALE* | -0.0009*** | | | -0.0009*** | | |
| | (-3.57) | | | (-4.06) | | |
| *GROUP* | | -0.0004 | | | -0.0003 | |
| | | (-1.49) | | | (-1.51) | |
| *MIX* | | | -0.0003 | | | -0.0003 |
| | | | (-0.56) | | | (-0.91) |
| *EDU* | -0.0006** | -0.0006** | -0.0006** | -0.0006*** | -0.0005*** | -0.0006*** |
| | (-2.36) | (-2.25) | (-2.42) | (-2.64) | (-2.55) | (-2.69) |
| *STAR* | 0.0009** | 0.0009** | 0.0009** | 0.0008*** | 0.0008*** | 0.0008*** |
| | (2.49) | (2.65) | (2.54) | (2.98) | (3.08) | (3.07) |
| *TRSIZE* | 0.0002*** | 0.0003*** | 0.0003*** | 0.0002*** | 0.0002*** | 0.0002*** |
| | (3.15) | (3.43) | (3.21) | (3.04) | (3.30) | (3.15) |
| *PRIOR_CAR* | -0.0051*** | -0.0052*** | -0.0051*** | -0.0047*** | -0.0048*** | -0.0047*** |
| | (-3.75) | (-3.79) | (-3.75) | (-5.66) | (-5.70) | (-5.66) |
| *BM* | 0.0021*** | 0.0021*** | 0.0021*** | 0.0019*** | 0.0019*** | 0.0019*** |
| | (5.44) | (5.21) | (5.44) | (5.76) | (5.83) | (5.76) |
| *SIZE* | 0.0038*** | 0.0037*** | 0.0037*** | 0.0036*** | 0.0036*** | 0.0036*** |
| | (17.41) | (17.40) | (17.41) | (20.91) | (20.90) | (20.91) |
| *ROA* | -0.0001 | -0.0005 | -0.0003 | 0.0007 | 0.0004 | 0.0005 |
| | (-0.02) | (-0.09) | (-0.06) | (0.15) | (0.09) | (0.12) |
| *Cons.* | -0.0570*** | -0.0572*** | -0.0574*** | -0.0518*** | -0.0520*** | -0.0522*** |
| | (-12.40) | (-12.46) | (-12.43) | (-14.24) | (-14.32) | (-14.29) |
| *Adj. $R^2$* | 17.87% | 17.88% | 17.85% | 30.74% | 30.71% | 30.70% |
| *Obs.* | 51,804 | 52,039 | 51,804 | 51,804 | 52,039 | 51,804 |

***, ** and * indicate significance level at 1%, 5% and 10%, respectively;

All estimations have controlled Year and Industry fixed effect;

Standard Errors are estimated by two-way cluster at firm and analyst level.

Results are estimated by restricting the number of sentences in each analyst report, *LENGTH,* being greater than or equal to 10; results are similar if not posing such restriction.

To test hypothesis described in H2a that analysts with welcomed (unwelcomed) characteristics yield lower (higher) forecast error, we estimate Equation (4) both for forecast error (*FE*) at **Table 5 Panel B** Column (1), (2), (3) and absolute forecast error (*AFE*) at **Table 5 Panel B** Column (4), (5), (6) separately for different team-level measurements. The interaction term of *OPN* and *SITE* is our first interest. We find positive and significant (P<0.01) correlation between the interaction term and absolute forecast error (*AFE)*, and the forecast error (*FE*) tend to increase with *OPN* after site visits, indicating post site visit reports are generally optimistic biased. One standard deviation increase of *OPN* after site visit would increase absolute forecast error by 5 base points per unit currency per share. This reinforces our previous assumption that analysts only conduct site visits to those firms they *believe* will outperform in recent. Secondly, we find some analyst extrinsic traits also significantly explains forecast accuracy. For (1) Female analysts being more conservative than male and (2) female analysts needing to work harder to gain self-identity in this male-dominating world, we observe have 9 base points lower absolute forecast errors (*AFE*). Nevertheless, females' exceptional effect would disappear if they work as a team with at least one male analyst. Education level is proved to be advantageous in analyst universe, master (doctor) are 5 to 6 base points more accurate than undergraduate (master) degree obtainers. Ironically, Star analysts are not masters for prediction accuracy – tending to be over-optimistic for 8 base points. Brokerage house size (*TRSIZE*) also works negatively to absolute prediction errors (*AFE*). This could be attributed to personal analysts' career concerns and social networks with firms' management teams outweighing the value of resources in those large brokerages.

Overall, we find that analyst gender (*FEMALE*) and education level (*EDU*) are normal welcomed characteristics, lowering prediction errors; yet conventionally agreed characteristics such as Star analysts (*STAR*), site visit (*SITE*), trading house size (*TRSIZE*) tend to work negatively and to boost their forecast errors.

## 6.3 Analyst Characteristics and Analyst Career Path Outcome

Having demonstrated that analyst characteristics could influence corresponding textual writing and prediction accuracy, we next focus on analyst characteristics with their career outcomes in our final set of analysis. In U.S., Star analysts earn 61% higher than

non-star analyst (Groysberg, Healy, and Maber, 2011); in China, though detailed salary reports are available, Star analysts typically enjoy higher wages and compensations, brighter occupation opportunities, and greater publicity (Li, 2017)

Two pivotal reasons justify our vanilla model initially set at report level instead of at yearly average level: (1) Star analysts are announced with 3 to 5 outstanding, far-influencing and representing analyst reports in *New Fortune* Magazine each year, suggesting that exceptional individual reports should be put on more weight and not be equally treated with others "inferiors"; (2) institutional investors, holding large percent of votes, also tend to veto those analysts with reports leading to great loss. Two examples are analysts recommending LeTV before its financial distress and Bitcoins before government harsh restrictions may cause great losses for fund managers.

By merging 281,884 full analyst reports from 2006 to 2016 reports with Star analysts list on yearly bases, we find 13.78% of them are from Star analysts. We design our Probit regression model as stated in Equation (5) to explore the probability of becoming a Star analyst using absolute forecast accuracy (*AFE*) and two-day cumulative returns (*CAR0_1*) similar analyst intrinsic and extrinsic variables in previous models. Since Star analysts are publicized in the end of November of each year, serving as a summary of yearly analyst universe, reverse causality problem is trivial here. The Probit model takes the following form:

$$\begin{aligned} P(Star = 1) \; &= \alpha_0 + \beta_1 APPRO + \beta_2 YOAT + \beta_3 DOM + \gamma_1 AFE + \gamma_2 CAR0\_1 \\ &+ \gamma_3 SITE + \gamma_4 EXPER + \gamma_5 FEMALE + \gamma_6 MIX + \gamma_7 TRSIZE \\ &+ \sum_j \delta_j Controls_j + \varepsilon \qquad (5) \end{aligned}$$

**Table 6** reports estimation results of Equation (5), with column (1) reporting coefficients of estimations without textual characteristics; and column (3) reporting corresponding average marginal effects. At report level, prediction accuracy, but not financial market return (*CAR0_1*) could predict career success likelihood, with one standard deviation decrease of absolute forecast error (*AFE*) increases successful probability by 0.78%. Frequent site visits decrease analysts' likelihood to be voted as Star analyst. This is reasonable since analysts with limited time often face the tradeoff between conducting site visitation and conducting roadshows to institutional investors. Those fund

investors, however, hold election votes for Star analysts in according to their fund size, but target firm managers normally do not.

**Table 6: Analyst Characteristics and Career Path**

**Panel A: Estimation of Analyst Characteristics and Career Path at Document Level**

| | (1) Star Coefficients (z-stats) | (2) Star Coefficients (z-stats) | (1) Star dy/dx (z-stats) | (2) Star dy/dx (z-stats) |
|---|---|---|---|---|
| *APPRO* | -0.0054*** | -0.0059*** | -0.0013*** | -0.0014*** |
| | (-3.59) | (-3.91) | (-3.59) | (-3.92) |
| *YO_AT* | 0.5396*** | 0.5436*** | 0.1275*** | 0.1271*** |
| | (15.78) | (15.69) | (15.86) | (15.77) |
| *DOM* | 0.4666*** | 0.4554*** | 0.1103*** | 0.1065*** |
| | (8.42) | (8.14) | (8.43) | (8.15) |
| *AFE* | -1.6281** | -1.6383** | -0.3848** | -0.3830** |
| | (-2.17) | (-2.17) | (-2.17) | (-2.18) |
| *CAR0_1* | -0.2069 | -0.2733 | -0.0489 | -0.0639 |
| | (-0.71) | (-0.93) | (-0.71) | (-0.93) |
| *SITE* | -0.1542** | -0.1651** | -0.0365** | -0.0386** |
| | (-2.32) | (-2.52) | (-2.33) | (-2.52) |
| *EXPER* | 0.0414*** | 0.0524*** | 0.0098*** | 0.0122*** |
| | (3.41) | (4.28) | (3.40) | (4.26) |
| *FEMALE* | -1.1880*** | -1.1366*** | -0.2808*** | -0.2657*** |
| | (-22.91) | (-22.06) | (-24.74) | (-23.71) |
| *MIX* | 0.9584*** | 0.9412*** | 0.2265*** | 0.2200*** |
| | (16.31) | (16.07) | (16.92) | (16.68) |
| *TRSIZE* | 0.9376*** | 0.9594*** | 0.2216*** | 0.2243*** |
| | (26.63) | (26.19) | (32.14) | (31.87) |
| *FIN* | | -0.0092 | | -0.0021 |
| | | (-1.26) | | (-1.26) |
| *COMPLEX* | | 0.0062*** | | 0.0015*** |
| | | (6.65) | | (6.73) |
| *CONFI* | | 0.0344*** | | 0.0080*** |
| | | (3.79) | | (3.80) |
| *TITLE* | | 0.0342* | | 0.0080* |
| | | (1.79) | | (1.79) |
| *CONCISE* | | 0.0010 | | 0.0002 |
| | | (0.52) | | (0.52) |
| *Cons.* | -25.5390*** | -26.4152*** | | |
| | (-30.48) | (-30.15) | | |
| *Pseudo-R2* | 35.04% | 35.89% | | |
| *Obs.* | 11,093 | 11,093 | 11,093 | 11,093 |

***, ** and * indicate significance level at 1%, 5% and 10%, respectively;
All estimations have controlled Year and Industry fixed effect;
Standard Errors are estimated by two-way cluster at firm and analyst level.

At individual level, we find analyst personality traits significantly (P<0.01) predict the probability of becoming a Star analyst: one standard deviation increases of Youthfulness and Attractiveness (*YOAT*) score increase probability by 7.50% and one standard deviation increase of Dominance (*DOM*) score increase probability by 3.78%. Seniority also exists in analyst field – one additional year of working experience as analysts (*EXPER*) increases about 1% likelihood of being voted as Star analyst. Currently, this field is male dominant, and male analysts are astonishingly 28.08% more likely to succeed than female (*FEMALE*) analysts. Team working seems to be promoted during elections and observed 22.65% greater chance of being successful for analyst team with at least one male and one female (*MIX*). Brokerage house size also plays a role during this process: analysts affiliated with larger, influential, and resources abundant trading houses are more likely to be elected as Star analyst.

Additionally, we also incorporate analyst report textual characteristics into our model to study whether certain textual traits could predict career success. Column (2) and (4) in **Table 6** reports estimation results coefficient and average marginal effect for this expanded model. Report confidence (*CONFI*) and complexity (*COMPLEX*) both positively predict favorable career outcome – one standard deviation increases of confidence and complexity level increase favorable outcome probability by 1.32% and 3.49%.

We next synthesize our report level data into yearly average data to examine the relationship between analysts yearly overall performance and probability of being selected as Star analyst. We shrink our sample by restricting an analyst or write 3 or more analyst reports, resulting in a total of 10,533 observations with 1,678 with identifiable profile images. The revised Probit model takes the following form:

$$\begin{aligned} P(Star = 1) &= \alpha_0 + \beta_1 APPRO + \beta_2 YOAT + \beta_3 DOM + \gamma_1 EXPER + \gamma_2 FEMALE \\ &\quad + \gamma_3 MIX + \gamma_4 TRSIZE + \gamma_5 NUMFIRMS + \theta_1 SIZE\_y + \theta_2 AFE\_y \\ &\quad + \theta_3 CAR\_y + \sum_j \delta_j Controls_j + \varepsilon \qquad (6) \end{aligned}$$

Four variables are computed to evaluate analyst yearly performance and analyst working patterns: yearly average of absolute forecast error (*AFE_y*), yearly average of

cumulative abnormal return (*CAR_y*), number of firms an analyst follows at a given year (*NUMFIRMS*) and the average size of firms followed (*SIZE_y*) following Bradley et al. (2017). **Table 7 Panel A** provides summary statistics of the additional yearly variable. On average, an analyst or a team of analysts follow 7 firms and rolls out 14 analyst reports, yielding significant upper biased forecasted EPS (*FEPS_y*) and providing significant yearly average two-day abnormal returns (*CAR_y*), with t-stats being 102.95 and 12.41, respectively. Industry fixed effects are ambiguous when data is grouped at year level since analyst may focus on one specific industry during one year or may shift between industries for multiple expediency reasons, thus we report estimation results with and without industry fixed effect separately.

**TABLE 6: Analyst Characteristics and Career Path (Cont.)**

**Panel B: Summary of Yearly Average of Analyst Reports Related Information**

| | Obs. | Mean | Std. Dev. | Min | Median | Max |
|---|---|---|---|---|---|---|
| *REPORT_NUM* | 10533 | 13.9350 | 14.1626 | 3.0000 | 9.0000 | 74.0000 |
| *N_FIRMS* | 10533 | 6.7835 | 9.0988 | 0.0000 | 4.0000 | 45.0000 |
| *N_INDUS* | 10533 | 6.3845 | 9.1518 | 0.0000 | 3.0000 | 45.0000 |
| *CAR_P2_P1_y* | 10533 | 0.0008 | 0.0080 | -0.0205 | 0.0000 | 0.0263 |
| *CAR0_1_y* | 10533 | 0.0009 | 0.0141 | -0.0209 | 0.0000 | 0.0305 |
| *CAR0_5_y* | 10533 | 0.0009 | 0.0163 | -0.0292 | 0.0000 | 0.0391 |
| *FIN_y* | 10533 | 0.3792 | 0.4379 | 0.0114 | 0.2276 | 2.0267 |
| *COMPLEX_y* | 10533 | 3.4844 | 4.3340 | 0.2517 | 2.3623 | 18.6599 |
| *CONFI_y* | 10533 | 0.3070 | 0.3393 | 0.0000 | 0.1888 | 1.5924 |
| *TITLE_y* | 10533 | 0.0755 | 0.1421 | -0.3333 | 0.0625 | 0.3333 |
| *CONCISE_y* | 10533 | -0.9781 | 1.5893 | -7.1255 | -0.4187 | 0.0000 |
| *SIZE_y* | 10533 | 2.7093 | 2.0150 | 0.0000 | 2.1469 | 7.5204 |

**Panel C: Estimation of Analyst Characteristics and Career Path at Yearly Level**

| | (1) Star Coefficients (z-stats) | (2) Star Coefficients (z-stats) | (3) Star Coefficients (z-stats) | (4) Star Coefficients (z-stats) | (1) Star dy/dx (z-stats) | (2) Star dy/dx (z-stats) | (3) Star dy/dx (z-stats) | (4) Star dy/dx (z-stats) |
|---|---|---|---|---|---|---|---|---|
| *APPRO* | -0.0052 | -0.0040 | -0.0043 | -0.0045 | -0.0011 | -0.0009 | -0.0009 | -0.0009 |
| | (-1.04) | (-0.80) | (-0.73) | (-0.76) | (-1.04) | (-0.80) | (-0.73) | (-0.76) |
| *YOAT* | 0.3796*** | 0.3471*** | 0.4952*** | 0.4642*** | 0.0836*** | 0.0752*** | 0.1007*** | 0.0928*** |
| | (3.98) | (3.51) | (4.07) | (3.71) | (3.98) | (3.50) | (4.14) | (3.74) |
| *DOM* | 0.4191** | 0.3648** | 0.3145 | 0.2627 | 0.0923** | 0.0791** | 0.0639 | 0.0525 |
| | (2.32) | (2.00) | (1.56) | (1.28) | (2.31) | (1.99) | (1.56) | (1.28) |
| *EXPER* | 0.1324*** | 0.1330*** | 0.1644*** | 0.1716*** | 0.0292*** | 0.0288*** | 0.0334*** | 0.0343*** |
| | (2.85) | (2.85) | (3.13) | (3.22) | (2.85) | (2.85) | (3.15) | (3.24) |
| *FEMALE* | -0.5552*** | -0.6155*** | -0.8811*** | -0.9041*** | -0.1223*** | -0.1334*** | -0.1791*** | -0.1807*** |
| | (-3.59) | (-3.92) | (-4.98) | (-5.02) | (-3.66) | (-4.01) | (-5.11) | (-5.15) |
| *MIX* | 0.8731*** | 0.9145*** | 1.0423*** | 1.0672*** | 0.1923*** | 0.1982*** | 0.2119*** | 0.2133*** |
| | (5.05) | (5.17) | (5.32) | (5.29) | (5.27) | (5.44) | (5.58) | (5.59) |
| *TRSIZE* | 0.6813*** | 0.7098*** | 0.7435*** | 0.7772*** | 0.1501*** | 0.1539*** | 0.1511*** | 0.1553*** |
| | (8.01) | (8.11) | (8.12) | (8.18) | (8.80) | (9.06) | (9.02) | (9.22) |
| *NUM_FIRMS* | 0.0192*** | 0.0191*** | 0.0193*** | 0.0196*** | 0.0042*** | 0.0041*** | 0.0039*** | 0.0039*** |
| | (4.64) | (4.30) | (4.43) | (4.10) | (4.79) | (4.41) | (4.55) | (4.20) |
| *SIZE_y* | -0.0252 | 0.0052 | -0.0259 | -0.0139 | -0.0055 | 0.0011 | -0.0053 | -0.0028 |
| | (-0.74) | (0.10) | (-0.72) | -0.24) | (-0.74) | (0.10) | (-0.72) | (-0.24) |
| *AFE_y* | | -10.7039 | | -15.5732* | | -2.3203 | | -3.1122* |
| | | (-1.46) | | (-1.84) | | (-1.47) | | (-1.86) |
| *CAR0_1_y* | | -9.7934 | | -10.6467 | | -2.1229 | | -2.1277 |
| | | (-1.18) | | (-1.24) | | (-1.18) | | (-1.25) |
| *FIN_y* | | 0.3215* | | 0.2381 | | 0.0697* | | 0.0476 |
| | | (1.83) | | (1.26) | | (1.83) | | (1.26) |
| *COMPLEX_y* | | -0.0055 | | -0.0025 | | -0.0012 | | -0.0005 |
| | | (-0.31) | | (-0.14) | | (-0.31) | | (-0.14) |
| *CONFI_y* | | -0.1850 | | -0.0538 | | -0.0401 | | -0.0108 |
| | | (-0.69) | | (-0.20) | | (-0.69) | | (-0.20) |
| *TITLE_y* | | -0.6548 | | -0.7338 | | -0.1419 | | -0.1466 |
| | | (-1.13) | | (-1.20) | | (-1.14) | | (-1.20) |
| *CONCISE_y* | | 0.0616 | | 0.0277 | | 0.0134 | | 0.0055 |
| | | (1.41) | | (0.60) | | (1.42) | | (0.60) |
| *SITE_y* | | 0.0126 | | -0.0074 | | 0.0027 | | -0.0015 |
| | | (0.32) | | (-0.17) | | (0.32) | | (-0.17) |
| *Year FE* | Controlled | Controlled | Controlled | Controlled | | | | |
| *Industry FE* | N | N | Controlled | Controlled | | | | |
| *Cons.* | -19.285*** | -19.868*** | -20.578*** | -21.297*** | | | | |
| | (-8.95) | (-8.99) | (-9.16) | (-9.12) | | | | |
| *Pseudo-R2* | 25.12% | 26.43% | 32.25% | 33.51% | | | | |
| *Obs.* | 803 | 803 | 775 | 775 | 803 | 803 | 775 | 775 |

***, ** and * indicate significance level at 1%, 5% and 10%, respectively;
Robust Standard Errors are estimated and reported.

**Panel B** of **Table 7** reports estimation results of Equation (6). We still find significantly (P<0.01) and positive coefficient of Youthfulness and Attractiveness score (*YOAT*), which reinforces our previous report level test. Analyst experience (*EXPER*) and group working (*MIX*) are also positively predicting successful likelihood at strong significance level (P<0.01); Female analysts faces more obstacles to become successful than male. Jointly, on year average level, we find homogeneous results as in Table 6, documenting our hypothesis H3a that analysts with welcomed (unwelcomed) personality traits are more (less) likely to become a Star analyst. Yearly summary variables also provide some distinct perspectives: more firms an analyst follows, more hardworking the analyst may be, thus higher the probability to be elected as a Star analyst; prediction accuracy over one year (*AFE_y*) and market cumulative abnormal returns over one year (*CAR_y*) are not significant, which supports our previously discussed functions of distinction reports. Previously developed textual information is not informative in this model.

Altogether, our findings support hypothesis H3a that an analyst has greater probabilities to become a Star analyst when one is young and stunning, when one is experienced and competent, when one cherishes teamwork, when one is male, and when one is industrious.

# 7. Conclusions

The emphasis that financial market participants address on sell-side analysts per se, in addition to their sideline behaviors, profoundly exceed existing efforts researchers has made to understand them. For lacking measurements of innate characteristic, little empirical evidence suggest the potential links between analysts' characteristics with their report content and performance, despite strong anecdotal observations. This paper attempts to fill this gap. By employing Naïve Bayes algorithm, we extract and analyze unstructured 1,068,679 analyst reports from 2006 to 2016; by employing key facial landmarks detection from standard profile images, we obtain 117,100 analysts' three

important personality traits from 62 developed facial attributes, providing the basis upon which we could bridge analyst characteristics and their general performance.

We find that analysts' characteristics are concealed and reflected in their analyst reports. Certain characteristics, both intrinsic and extrinsic, could predict general opinion level of textual content and, further, specific textual characteristics. Moreover, we extend our study to financial market and document that analysts' extrinsic characteristics are associated with their prediction accuracy – welcomed (unwelcomed) characteristics may bring them more (less) private information and subsequently lower (higher) observed prediction errors. Site-visitation, specifically, is negatively related to analyst performance for it letting analysts being more optimistic biased and lowing corresponding prediction accuracy. When we examine analyst characteristics and their career outcomes, we find that analysts with welcomed characteristics are more likely to have favorable career outcome.

Finally, we document the effectiveness of Naïve Bayes algorithm to Chinese-language based content in China financial market, and the effectiveness of personality traits extraction approach from standard profile images by analyzing key facial landmarks and underlying facial attributes. The novel approach which combining both natural language processing (NLP), psychology and computer science, provide a novel perspective to understand the analyst universe.

# References


[1] Allen H. Huang, Amy Y. Zang, Rong Zeng, 2014, Evidence on the Information Content of Text in Analyst Reports, The Accounting Review, [J], Vol. 89, No. 6, pp. 2151–2180

[2] Allen H. Huang, Reuven Lehavy, Amy Zang, Rong Zheng, 2016, Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach, Working Paper, No.1229, available at http://ssrn.com/abstract=2409482

[3] Athey, S., C.Avery, and P. Zemsky, 2000, Mentoring and discrimination, American Economic Review 90:765–86.

[4] C. Sagonas, E. Antonakos, G, Tzimiropoulos, S. Zafeiriou, M. Pantic, 2016, 300 faces In-the-wild challenge: Database and results. Image and Vision Computing (IMAVIS), Special Issue on Facial Landmark Localization, "In-The-Wild"

[5] Daniel Bradley, Sinan Gokkaya, Xi Liu, 2017, Before an Analyst Becomes an Analyst: Does Industry Experience Matter? [J], The Journal of Finance, Vol. LXXII, No. 2

[6] Fei-fei Li, Justin Johnson, Serena Yeung, 2017, Computer Science 231n, Lecture Slides 7, Stanford University, available at http://cs231n.stanford.edu

[7] Feng Li, 2011, Textual Analysis of Corporate Disclosures: A Survey of the Literature, [J], Journal of Accounting Literature (Forthcoming)

[8] Groysberg, Boris, Paul M. Healy, and David A. Maber, 2011, What Drives Sell-Side Analyst Compensation at High-Status Investment Banks? Journal of Accounting Research 49, pp.969 –pp.1000

[9] Groysberg, Boris, Paul M. Healy, and David A. Maber, 2011, What drives sell-side analyst compensation at high-status investment banks? Journal of Accounting Research 49, 969–1000.

[10] Harrison Hong and Jeffrey D. Kubik, 2003, Analyzing the Analysts: Career Concerns and Biased Earnings Forecasts, [J], The Journal of Finance, Vol. LVIII, No.1

[11] Jiawen Yan, Yakun Wang, Yirong Kang, 2017, Evidence of Information Content of Text in China's Listed Company's Notices, Working Paper

[12] Joseph D. Piotroskia, T.J. Wong, Tianyu Zhang, 2016, Political Bias of Corporate News: Role of Conglomeration Reform in China, Working Paper

[13] Jungseock Joo, Francis F. Steen, and Song-Chun Zhu, 2017, Automated Facial Trait Judgment and Election Outcome Prediction: Social Dimensions of Face, Working Paper, UCLA

[14] Lewis, D. 1998. "Naive (Bayes) at forty: The Independence Assumption in Information Retrieval", Proceedings of CEML-98, 10th European Conference on Machine Learning, 4-15

[15] Li Fang, 2017, New Fortune Magazine, Vol.12 ISSN 1671-1319

[16] Li, F, 2010, The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach, [J], Journal of Accounting Research. Vol. 48 No. 5 December:1049-102.

[17] Lily Hua Fang, Sterling Huang, 2017, Gender and Connections among Wall Street Analysts, [J], The Review of Financial Studies, Vol. 30 No. 9

[18] Manning, C. D., and H. Schütz. 1999, Foundations of Statistical Natural Language Processing, Cambridge, MA: MIT Press.

[19] Michael Firth, Chen Lin, Ping Liu, and Yuhai Xuan, 2013, The Client Is King: Do Mutual Fund Relationships Bias Analyst Recommendations? [J], Journal of Accounting Research, Vol.51 No.1

[20] Pang, B., and L. Lee. 2008. Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval (2): 1-135.

[21] Pilar Corredor, Elena Ferrer, Rafael Santamaria, 2014, Is cognitive bias really present in analyst forecasts? The role of investor sentiment, [J], International Business Review, Vol.23 pp.824 – pp.837

[22] Richard J. W. Vernon, Clare A. M. Sutherland, Andrew W. Young, and Tom Hartley, 2015, Modeling First Impressions from Highly Variable Facial Images, 2014, Proceedings of the National Academy of Sciences of the United States of America, [J], E3353–E3361

[23] Shuping Chen, Dawn Matsumoto, Shiva Rajgopal, 2011, Is Silence Golden? An Empirical Analysis of Firms That Stop Giving Quarterly Earnings Guidance, [J], Journal of Accounting and Economics, Vol.51, pp134 –pp.150

[24] Simona Mola and Massimo Guidolin, 2009, Affiliated Mutual Funds and Analyst Optimism, Journal of Financial Economics, [J], Vol. 93, pp.108 – pp.137

[25] Ting Zhang, Ri-Zhen Qin, Qiu-Lei Dong, Wei Gao, Hua-Rong Xu, Zhan-Yi Hu, 2017, Physiognomy: Personality Traits Prediction by Learning, [J], International Journal of Automation and Computing

[26] Twedt, B., and L. Rees. 2012. Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports. Journal of Accounting and Public Policy 31 (1): 1–21.

[27] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Boudev, 2012, Thomas S. Huang, Interactive Facial Feature Localization, ECCV 2012

[28] Wilma A. Bainbridge, Phillip Isola, Aude Oliva, 2013, The Intrinsic Memorability of Face Photographs, Journal of Experimental Psychology, Vol. 142, No. 4, 1323–1334

[29] Xia Chen, Qiang Cheng, Kin Lo, 2009, On the Relationship Between Analyst Reports and Corporate Disclosures: Exploring the Roles of Information Discovery and Interpretation, Journal of Accounting and Economics, [J], Vol.49, pp.206–pp.226

[30] Xianjie He, Huifang Yin, Yachang Zeng, Huai Zhang, 2016, Achievement Drive and Analysts' Performance, Working Paper, available at https://ssrn.com/abstract=2921712

[31] Yuping Jia, Laurence Van Lent, Yachang Zeng, 2014, [J], Journal of Accounting Research, Vol. 52 No. 5

# Appendix

## A1: Detailed Data Description

### A1.1 Textual Analyst Reports

Analyst reports are crawled from *Tencent Finance* and *East Money*, two major financial information providers in mainland China. For those reports in HTML (HyperText Markup Language) format, information such as analysts, date, stock code, title etc. are extracted by analyzing tagged elements; For those reports in PDF (Portable Document Format) format, PDF ToolBox, a Java API in employed to extract structured content. Noise information such as disclaimers, brokerages' introduction, analysts' introduction etc. are removed from main content.

**TABLE A1: Comparison of The Number of Recommendations in Each Category Between China and U.S. Analyst Reports**

| | CHINA | | | U.S. (Huang et al., 2014) | | |
|---|---|---|---|---|---|---|
| | **Freq.** | **Percent** | **Cum.** | **Freq.** | **Percent** | **Cum.** |
| **Recommendation Level** | | | | | | |
| Strong Buy | 77,761 | 49.97% | 49.97% | 177,883 | 55.32% | 55.32% |
| Buy | 66,570 | 42.78 | 92.75 | | | |
| Neutral | 10,902 | 7.01 | 99.75 | 123,304 | 38.35 | 93.67 |
| Sell | 133 | 0.09 | 99.84 | 20,346 | 6.33 | 100.00 |
| Strong Sell | 251 | 0.16 | 100.00 | | | |
| **Recommendation Revision** | | | | | | |
| Upward | 2,772 | 2.07% | 2.07% | 8,588 | 2.73% | 2.73% |
| Reiteration | 126,252 | 94.06 | 96.12 | 297,274 | 94.46 | 97.19 |
| Downward | 5,205 | 3.88 | 100.00 | 8,841 | 2.81 | 100.00 |

### A1.2 Analyst Profile Image

117,100 analysts profile from 129 brokerage houses are collected from SAC's (Security Association of China) website, www.sac.net.cn. Each Analyst's name, education, affiliated brokerage, gender as well as a 14-digits unique identification number are disclosed in profile page, allowing as to match with textual analyst reports.

### A1.3 Sample Data Description

Total analyst reports from Tencent Finance News, EastMoney.com: 1,068,679

Subtract industry level analyst reports and market level analyst reports: 309,056

Subtract analyst reports without Recommendation, Forecasted EPS or CAR: 165,820

(1) Merged with analyst characteristics excluding personality traits: 110,830

Further merged with analyst personality traits: 12,374

(2) Merged with SUE: 60,355

Restricting number of sentences per report greater than ten: 52,039

## Appendix 2: Modelling 68 Fiducial Points and Calculating Personality Traits Scores

### A2.1 Modelling 68 Fiducial Points

A 68 facial landmarks shape predictor model is trained using iBUG 300w dataset (see further information and discussion on dlib.net). With the pre-trained model and the Python wrapped API module, we could efficiently predict 68 key facial landmarks for each image over 200 thousand analysts profile image within 24 hours. Examples of predicted facial landmarks are shown as below.

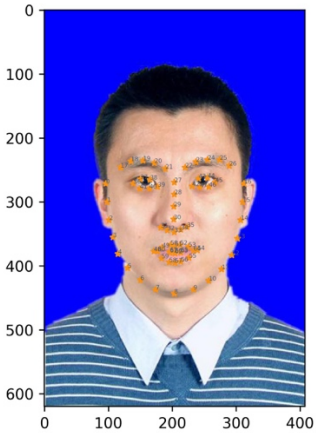



Fig. A.1

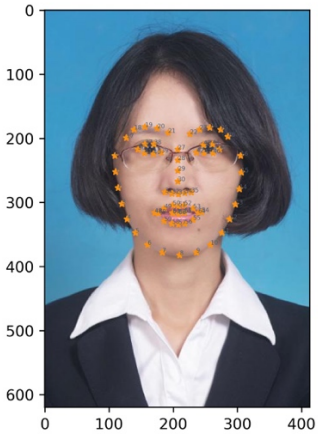



Fig. A.2

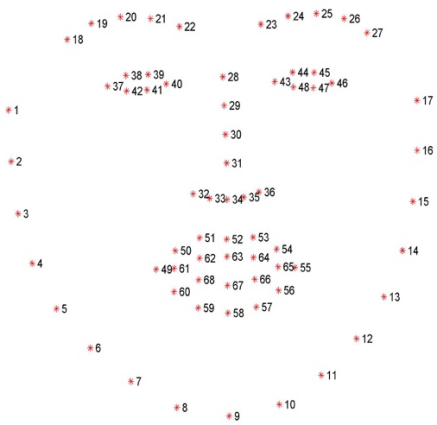


Fig. A.3

## A2.2 Calculating Personality Traits

Vernon et al. (2014) derived 3 personality traits, namely, approachability, youthfulness and attractiveness and dominance, from 65 facial attributes modelled by 179 fiducial points. Top 62 facial attributes are modelled by 68 key facial landmarks, leaving glasses (63), facial hair color (64) and stubble (65), which are not directly calculable.

Though only 68 points are available in our prediction model, they are enough to capture those identifiable personality attributes; and we remodeled some calculation in Vernon et al (2014) to suit our case. (See Vernon et al, 2014 for detailed derivation of 65 facial attributes)

Each Personality trait score is calculated as linear combination of highly significantly correlated ($P < 0.001$) facial attributes[1], with each trait value standardized by its mean. (Vernon et al, 2014)

Formally, each personality trait is calculated as follows:

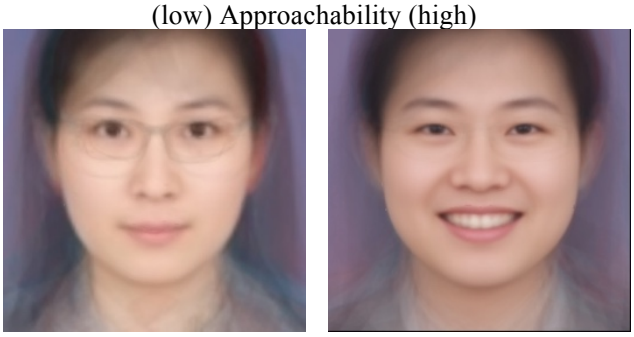
$$\begin{aligned} APPRO = {} & 0.14\times Attributes_3 + 0.17\times Attribuets_5 + 0.19\times Attributes_6 + 0.16\times Attributes_7 \\ & - 0.15\times Attributes_8 - 0.26\times Attributes_{11} - 0.20\times Attributes_{12} \\ & - 0.30\times Attributes_{13} - 0.31\times Attributes_{15} + 0.26\times Attributes_{16} \\ & + 0.45\times Attributes_{18} + 0.37\times Attributes_{19} - 0.37\times Attributes_{20} \\ & + 0.17\times Attributes_{21} + 0.18\times Attributes_{22} + 0.18\times Attributes_{24} \\ & + 0.69\times Attributes_{25} + 0.51\times Attributes_{26} - 0.24\times Attributes_{27} \\ & - 0.35\times Attributes_{28} + 0.73\times Attributes_{29} + 0.71\times Attributes_{30} \\ & + 0.36\times Attributes_{31} + 0.75\times Attributes_{32} + 0.22\times Attributes_{33} \\ & + 0.16\times Attributes_{34} - 0.23\times Attributes_{36} + 0.38\times Attributes_{39} \\ & - 0.39\times Attributes_{42} - 0.60\times Attributes_{47} - 0.24\times Attributes_{59} \end{aligned}$$

[1] In fact, we also tested facial attributes with significance level $P<0.05$ as a robustness check; results are similar

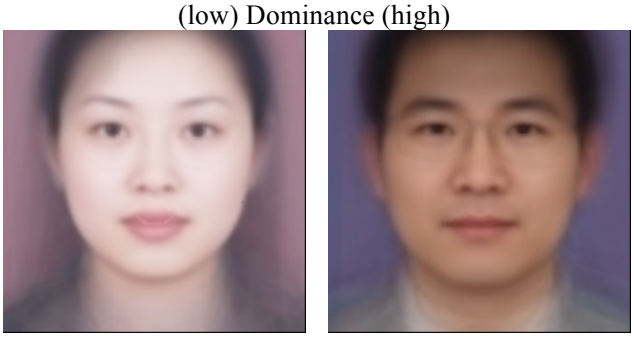
$$YOAT = 0.18\times Attributes_3 + 0.28\times Attribuets_5 + 0.20\times Attributes_6 - 0.21\times Attributes_7 + 0.33\times Attributes_8 + 0.22\times Attributes_9 + 0.31\times Attributes_{10} + 0.40\times Attributes_{11} + 0.41\times Attributes_{12} + 0.39\times Attributes_{13} + 0.34\times Attributes_{14} + 0.24\times Attributes_{15} + 0.24\times Attributes_{17} + 0.35\times Attributes_{21} + 0.33\times Attributes_{22} + 0.25\times Attributes_{23} + 0.31\times Attributes_{24} + 0.15\times Attributes_{26} + 0.24\times Attributes_{27} + 0.34\times Attributes_{28} - 0.17\times Attributes_{35} - 0.21\times Attributes_{36} - 0.28\times Attributes_{39} - 0.22\times Attributes_{40} + 0.23\times Attributes_{41} + 0.19\times Attributes_{42} - 0.38\times Attributes_{46} + 0.19\times Attributes_{55} - 0.21\times Attributes_{60} - 0.22\times Attributes_{61} - 0.24\times Attributes_{62}$$

$$DOM = -0.20\times Attributes_3 + 0.23\times Attributes_7 + 0.27\times Attributes_8 - 0.15\times Attributes_{10} - 0.22\times Attributes_{11} - 0.31\times Attributes_{12} - 0.19\times Attributes_{14} + 0.16\times Attributes_{18} + 0.14\times Attributes_{23} - 0.15\times Attributes_{25} - 0.22\times Attributes_{26} - 0.25\times Attributes_{27} - 0.15\times Attributes_{28} + 0.37\times Attributes_{35} + 0.32\times Attributes_{36} - 0.27\times Attributes_{38} - 0.21\times Attributes_{41} - 0.44\times Attributes_{43} + 0.28\times Attributes_{49} - 0.23\times Attributes_{50} + 0.15\times Attributes_{52} - 0.22\times Attributes_{53} + 0.21\times Attributes_{61} + 0.25\times Attributes_{62}$$

(low) Approachability (high)



(low) Dominance (high)



(low) Youthfulness and Attractiveness (high)

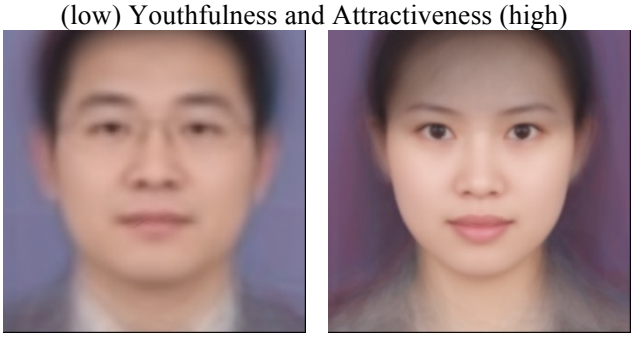


**Fig. A.4**

# 致谢

本篇论文的完成，结合了本科期间基础学科的学习、文献的阅读、自身计算机设计特长和导师的指导。在完成这篇论文过程中，我积极查找文献、主动思考、解决问题，收获颇多。在此，我要感谢许多人。

首先，我要真诚地感谢我的导师杜茜茜老师。从论文开题、数据处理、模型选择到论文定稿，杜老师都耐心仔细地教导我完成，并不时提出积极建设性意见。杜老师渊博的专业知识、严谨的治学态度、以及对中国证券分析师的独到见解使我受益良多。本篇论文的完成，离不开杜老师的悉心教导与培养。

其次，我要由衷地感谢经济与管理研究院所有老师对我的培养。大学四年丰富专业课的学习培养了我创新的思考方式、严谨的科研态度和坚实的基础知识。良好的金融学和经济学基础，这是我完成本篇论文的基础。

最后，我要衷心地感谢西南财经大学为我提供了良好的学习氛围。丰富的自由选修课程让我有机会尝试其他学院丰富的计算机和数学建模课程，在学习过程中积累的知识和良师益友不断鼓舞着我前进。

经世济民

致敬以求