Does the fluency of its name affect social media sentiment towards a company?

Tatenda Gatawa

Student Number 446154

A research project submitted to the Gordon Institute of Business Science, University of Pretoria, in partial fulfilment of the requirements for the degree of Master of Business Administration.

9 November 2015

ABSTRACT

Research Purpose: The recent change in name by Google to Alphabet sparked a worldwide interest. This research studies the extent to which stock sentiment data expressed online is influenced by the fluency of the corporates name. It is concerned with the extent to which seemingly banal factors that affect human sentiments create bias in opinions that are formed online. The research is carried out at a time in which the online environment and crowdsourcing are increasingly competing with traditional business models and platforms for primacy. • The purpose of this report is to quantify the extent to which crowd sourced stock sentiment data obtained from online sources is subject to bias resulting from the effect of spurious variables that are unrelated to the underlying fundamentals of the company. The biasing variable studied in this instance is the linguistic fluency of the corporates name.

Research Methodology: The study adopted a purely quantitative approach. Analysis was done using SAS

Research Findings: The research conducted makes a key finding that shows empirically that social media stock sentiment data are more influenced by the spurious effect of fluency of corporate name than are stock sentiments generated by traditional expert communities. This finding flies in the face of an increasingly adopted thesis that postulates superior intelligence can be obtained from an aggregation of unstructured crowd-sourced data. The research conducted also makes a key finding that the model of sentiment does not predict equally well at all levels of sentiment – models are better at identifying highly recommended equities than they are at identifying those that are not

KEYWORDS

Fluency

Sentiment

Social media

Cognitive bias

Moderating effects

DECLARATION

I declare that this research project is my own work. It is submitted in partial fulfilment of the requirements for the degree of Master of Business Administration at the Gordon Institute of Business Science, University of Pretoria. It has not been submitted before for any degree or examination in any other University. I further declare that I have obtained the necessary authorisation and consent to carry out this research.

…………………………………………………………………….

Tatenda Gatawa

9 November 2015

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction to Research Problem

1.1 Research title

Does the fluency of its name affect social media sentiment towards a company?

1.2 Background of study

On Tuesday the 10th of August 2015, investors worldwide were not only shocked but confused when they learnt that Google, one of the most famous and biggest company in the world, had announced its restructure into a new holding company called Alphabet (Hern, 2015). A fascinating article in the Guardian (Rushe & Sam, 2015) speculates as to the reason why Google , the citadel of citadels when it comes to the analysis of the most banal of data to provide an uplift in decision making, could possibly have elected to choose the name "Alphabet" for their new holding company. The article leaves one none the wiser and instead speculates that it may be a "mixture of terrible jokes, grand ambition, and carefully studied banality. The pun comes from the fact that "alpha" is a financial term meaning return on investment above the benchmark, making Alphabet a good Alpha-bet. In addition, the ambition comes from the fact that the Alphabet is one of humanity's most important inventions, as well as the implicit claim (in the company's abc.xyz url) that it encompasses everything from 'A' to 'Z' (Rushe & Sam, 2015). The banality comes from the fact that Alphabet is perhaps the most generic name imaginable, perfectly standing for anything and nothing at the same time. (Rushe & Sam, 2015)

Far from agreeing with the authors, the writer of this research is under no illusion that the scientists at Google have in this particular instance relaxed the rigour of their decisions and rather suspect that this is a powerful instance of the phenomenon studied in this research report.

This research studies the extent to which stock sentiment data expressed online is influenced by the fluency of the corporates name. It is concerned with the extent to which seemingly banal factors that affect human sentiments create bias in opinions that are formed online. The research is carried out at a time in which the online environment and crowdsourcing are increasingly competing with traditional business models and platforms for primacy.

The concept of fluency measures the ease with which individuals process external stimuli. Researchers in cognitive psychology and behavioural economics have shown that fluency has a positive effect on affinity or favourable judgement across a wide range of applications. For example, perceivers judge food additives with fluent names as less risky than those with less fluent names, believe currencies with fluent names to be more valuable and evaluate art that is processed more fluently more positively (for a review see Lick and Johnson, 2015).

1.3 Research problem

- The purpose of this report is to quantify the extent to which crowd sourced stock sentiment data obtained from online sources is subject to bias resulting from the effect of spurious variables that are unrelated to the underlying fundamentals of the company. The biasing variable studied in this instance is the linguistic fluency of the corporates name.

The first contribution it aims to make is to introduce a direct measure of sentiment into a model of corporate name fluency and thus perform rigorous statistical analysis to test if bias effects really exist online and if the extent of such bias differs from traditional platforms.

1.4 Research objectives

The research objectives are:

- To quantify the extent to which fluency moderates the relationship between stock fundamentals and the opinions expressed on such stocks online
- To quantify the extent to which these biases may differ when online platforms are compared to traditional sources of stock recommendations

The overarching objective is to improve the understanding of online stock data for investors that would like to harness this data in their research.

1.5 Report layout

The structure of the report will be a follows:

Chapter One – Chapter one introduces the back ground of the study and asks the question of whether the fluency of its name affect social media sentiment towards a company. In addition, this chapter discusses the motivation for the research, the problem of the research and the scope covered by the research.

Chapter Two – This chapter presents literature arguments which support why this research is relevant. Literature covered here includes some empirical and theoretical research already done in areas covering literature on relationships between linguistic fluency based aesthetics and the biasing effects of corporate names online and, secondly, literature which supports the proposition that developments in the understanding of fluency-based aesthetics have potential implications for users of online stock sentiment data

Chapter Three – This chapter introduces the four research questions after the background discussion on literature in chapter two.

Chapter Four – This chapter discusses the purely quantitative research methodology used in the research. A comparative analysis was done by analysing historical stock sentiment data and calculating the fluency of corporate name of each of the stocks considered in order to correlate the sentiment data to corporate name fluency.

Chapter Five – Chapter Five presents the findings of the data analysed on the investment analyst data and the social media sentiment data. Graphs and tables from SAS are also presented in this chapter.

Chapter Six – Chapter Six discusses the results presented in chapter five and how these results answer the research problems in chapter two and three.

Chapter Seven – This chapter concludes the research by looking at the research objectives and discussing the extent to which these research objectives were answered by the research findings. Recommendations which are based on literature and the findings in chapter five and six are also discussed. The chapter also discusses limitations and suggestions for future research.

Chapter Eight – All references used in the research are in this chapter.

Appendices – Appendices Consistency matrix and all additional information used in this research is presented here.

2 Literature Review

2.1 Introduction

The aim of this literature review is firstly, to establish the relationships between linguistic fluency based aesthetics and the biasing effects of corporate names online and, secondly, to propose that developments in the understanding of fluency-based aesthetics have potential implications for users of online stock sentiment data.

2.2 Drivers of the relationship between fluency and sentiment

The different mechanisms through which fluency could be related to sentiment have different implications for understanding how sentiment could mediate the relationship between a company's performance and the sentiment directed towards that company. These possible drivers therefore suggest contrasting practical strategies that should be employed to manage the effects of fluency on sentiment and for that reason they should be well understood.

This point is illustrated by looking at the example of two alternative mechanisms through which fluency could be defined and linked to sentiment, namely bounded rationality and aesthetic appeal (or attractiveness).

Fiske and Taylor (1991) showed that far from being fully rational operators, people are "cognitive misers" and often prefer to expend the minimum possible mental effort in a situation in order to arrive at an acceptable conclusion. This tendency includes the use of "mental shortcuts" such as stereotypes and other bases of snap judgement when making decisions and value judgements. This is very much aligned to the well-defined concept of bounded rationality which suggests that there are limitations to human rationality imposed by cognitive capacity, data and time pressure (Kahneman, 2003).

The concept of fluency measures the ease with which individual's process external stimuli. "Fluent processing is easy on the mind, marked by swift and seamless progress toward stimulus recognition and judgment. Disfluent processing is hard on the mind, marked by slow and effortful progress toward stimulus recognition and judgment" (Lick and Johnson, 2015).

Fluency therefore measures the amount of effort with which individual's process information bearing in mind that people are typically cognitive misers (Lick & Johnson, 2015). As such familiarity increases fluency and it has been shown in this context of bounded rationality that investors are attracted to more familiar i.e. fluent stocks and "overweight" them in portfolios (Green & Jame, 2013). In other instances researchers studied the effect of name fluency and demonstrated that easy-to-pronounce names elicited more affinity and people with fluent names held higher positions in law firms (Laham, Koval, & Alter, 2012). In a similar vein research showed that survey participants predicted superior financial performance for fictional stocks with more fluent names (Alter & Oppenheimer, 2006).

On the other hand aesthetic appeal could be an alternative and important measure of fluency which is not necessarily linked to ease of processing - consider for example the case of abstract art (Graf & Landwehr, 2015).

The aesthetic appeal of a company has been shown to be significantly affected by design elements used in its logo such as the typeface, colour, proportionality, and the 'favourability' of its name significantly affecting perceptions of modernity and innovativeness (Foroudi, Melewar, & Gupta, 2014) . On the other hand, the familiarity of a company's name can be managed through brand awareness initiatives such as sponsorships and advertisements and an understanding of the factors that increase the effectiveness of brand association (Olson & Thjømøe, 2009). Depending on the mechanism through which fluency is shown to most affect sentiment, executives would have a choice of different interventions to prioritise in order to manage sentiment.

The examples show how the different measures of fluency have significant implications for management. The effect of these alternative measures can be explored using the measures of fluency discussed below.

However it should be borne in mind that the separation between the different measures of fluency may in fact be minimal or they may be highly correlated. For instance it has also been shown that purely design and aesthetic elements such as the colour, proportionality and typeface of its logo can elicit feelings of familiarity with a company (Foroudi et al., 2014).

The discussion above underscores the necessity of considering the drivers of the relationship between fluency and sentiment when completing our current research on the effects of fluency on social media sentiment towards a company. These drivers and their measures are discussed further below.

2.3 Definition of key constructs

The sections below discuss the key concepts referred to in this research report.

2.3.1 Social media

Whilst the effect of fluency on sentiment has been studied in the context of expert investors, research panels and based on a broad number of proxies for sentiment, the environment of social media has significant different and unique characteristics that warrant a separate investigation of the effect of fluency in the context of social media.

The effects of fluency of corporate name are increasingly subject of research in recent times. Existing research (Green and Jame, 2014) investigates the impact of fluency on a company's fundamentals (specifically breadth of ownership, liquidity and firm value) and theorises that this effect is mediated by investor sentiment and investor recognition without including direct measures of sentiment in the model (i.e. proxy measures such as the breadth of ownership and stock liquidity are applied). Srinivasan and Umashankar (2014) studied the interaction between the fluency of a corporates ticker symbol and other elements of performance such as marketing and distribution in determining the 'intangible value' of a corporate and found significant interactions. Similarly Lee and Baack investigated the primacy of meaning or brand name fluency in terms of effect on brand recall and proclivity to purchase and once again established that fluent brand names had significant effects on both outcomes.

Whilst these studies provide a solid grounding for hypothesising a relationship between corporate name fluency and firm performance, the context of social media provides opportunities to explore this relationship and confirm elements that are not addressed in the current research.

The first opportunity presented by social media platforms is the burgeoning availability of large volumes of direct measures of sentiment that are being generated by machine learning algorithms that can trawl through social media content and score it in terms of sentiment expressed in the data (Cambria & White, 2014; Manke & Shivale, 2015; Pang &

Lee, 2008). Data generated by natural language processing algorithms far from being untested and possibly unreliable are being successfully applied in a wide range of contexts including machine translation, counter-terrorism and automated stock trading algorithms (Hirschberg & Manning, 2015).

The extent of relationship between fluency and sentiment may differ when small populations (such as panels of experts and other relatively smaller populations) are compared to the large volume of contributors available online as a result of a phenomenon commonly referred to as ‘the wisdom of crowds’. The wisdom of the crowds postulates that more accurate information is obtained from averaging large and diverse volumes of opinion as opposed to consulting experts. This has interesting implications for applications to stock picking as an example. The availability of social media sentiment data provide an opportunity to compare the fluency effect in these different contexts.

The importance of such a refined understanding cannot be overstated in the current context in which high-frequency trading algorithms that use sentiment analysis of news sources collected digitally and in real time as input has been shown to have the capacity to cause considerable adverse effects including the ability to trigger market panics (Kleinnijenhuis, Schultz, Oegema, & van Atteveldt, 2013) . Such algorithms are increasingly looking to the social media base as additional source of data input of automated sentiment data (Nardo, Petracco-Giudici, & Naltsidis, 2015)

These considerations underscore the relevance of a study of the effect of fluency of corporate name on sentiment in social media settings

2.3.2 Sentiment

Despite the long standing interest and accumulation of research into the effects of sentiment in financial markets, the analysis of sentiment in the context of financial markets is still the subject of serious analysis and important debate and therefore any study that can contribute to understanding the formulation of sentiment and the existence is relevant and can provide input actual applications.

The major reason for this is that it is not only sufficient to understand and quantify the existence of sentiment – understanding the causes of sentiment which may create cognitive biases in important drivers of market performance and investment strategy such as investment analysts and increasingly alternative sources of information such as social media has implications for the remedies that have to be employed to correct market biases.

Thus recent studies have investigated various possible sources of cognitive bias and optimism in financial markets including tactical intent of analysts, investor sentiment and emotion, stock characteristics (rarity or ease of acquisition), extent of analyst of coverage, national character and corporate governance regimes (for a review see Corredor, Ferrer, & Santamaria, 2014) and have re-fuelled a debate around the actual drivers of sentiment (i.e. cognitive bias or manipulation).

A comparative study of fluency effects on sentiment in social media settings can contribute to this understanding and therefore make a relevant and meaningful contribution.

2.3.2.1 Definition of the concept of sentiment

A fundamental difficulty with respect to studying sentiment is the subjective nature of the construct which presents challenges in measuring and reporting on it. In this context an investigation which refers to a previously unexplored measure of sentiment potentially adds value even if previously explored theoretical frameworks are applied.

Sentiment is a perception based construct which limits the extent to which it can be explored empirically. The most widely accepted empirical measures of investor sentiment are calculated at an overall market level and include measures such as the closed-end fund discount rate, share turnover, number of IPOs, first day return on IPOs, the market dividend premium and the equity share in new issues (Huang, Jiang, Tu, & Zhou, 2014).
The difficulty with market wide measures of investor sentiment is that they cannot be applied to investigate firm level influences on investor sentiment such as individual firm performance and fluency of corporate name.

Some of the firm level measures of sentiment that have been applied in recent research include changes in the breadth of ownership, stock liquidity and firm value (Green & Jame, 2013), analyst coverage (Hong & Kacperczyk, 2009), the price to earnings multiple and increasingly bespoke approaches such as the extent of inclusion in industry benchmarks and indices (Hacıbedel, 2014).

A number of recent research reports have also looked at social media sentiment and its possible effects on various firm level initiatives with applications mostly in the marketing and customer relationship management area (Kim, Koh, Cha, & Lee, 2015; Wei, Song, & Rutherford, 2014)

The various approaches and sources referred to in understanding sentiment have provided insight and underscore how this application which considers a twitter based data source will find relevance.

2.3.3 Themes from definition of sentiment

A number of salient themes arise out of the preceding definitions of sentiment most importantly:

- The consideration of type of source of sentiment data - sample or crowd, and
- The validity and necessity of testing multiple measures of sentiment

2.3.4 Fluency

The next construct considered in this section is fluency. This research explores a number of constructs of fluency that are suggested as valid measure of fluency by previous research findings. As discussed in the preceding sections fluency of a corporate name may also be viewed from a variety of perspectives which have different implications for the strategies that management could employ to address its implications for sentiment and overall company performance. As such it is necessary to include multiple measures of fluency.

2.3.4.1 Measures of fluency

In general fluency describes the subjective ease with which people process mental stimuli or information. This research is concerned with linguistic fluency which measures the phonological simplicity (pronunciation) and lexical simplicity (difficulty and awareness) of textual information (Green and Jame, 2014). Whilst this leans towards theories that contextualise the relationship between fluency and sentiment in terms of bounded rationality and cognitive biases, and attempt is made in this research to capture the potential effect of relationships between fluency and sentiment that may arise from aesthetic preferences of the population in alternate measures of fluency.

2.3.4.2 Corpus based measures of fluency

This research investigates linguistic fluency which is a measure of the ease with which individuals process linguistic stimuli. Ease of processing (or fluency) may come about as a result of familiarity.

Several official corpi exist for the English language which have been painstakingly compiled by linguistic experts. The Oxford English Corpus research programme for example maintains a database of more than 2 billion words drawn from varied sources such as newspapers, the universe of published books and specialty journals, websites, blogs and online chat rooms which (among other uses such as the dictionary research programme) is analysed to rank words in English by frequency of usage (Stevenson, 2010).

Another example is the Google Corpus which is a trillion word data set that was developed by Google and which includes a rich analysis of English language through its ability to capture all *tokens* used in the English universe (such as “LOL” and “.com” for example) in addition to ordinary dictionary words (Segaran & Hammerbacher, 2009). Whilst the English language has approximately one million dictionary words the Google Corpus has analysis on approximately 13 million natural language “tokens”.
A language corpus is a linguistic and quantitative analysis of all the words in a language based on current actual usage and provides a potential source of measuring both the familiarity and current preference for words (aesthetic preference) and therefore provides a valid source for analysing the fluency of corporate names.

2.3.4.3 Alphabeticity

A further theorised mechanism through which fluency can arise is through the ease through which information is acquired and process (Alter & Oppenheimer, 2006) . As a result of the commonly adopted practice of listing ticker symbols and most compilations of investor information that can be sourced online and in investment handbooks in alphabetical order it has been shown (consistent with the theory that humans can be cognitive misers and are not necessarily rational or thorough in their behaviour) firms with names and ticker symbols that occur higher in the alphabetical order are more widely purchased and traded out of the thousands of stocks available to be considered by potential investors (Itzkowitz, Itzkowitz, & Rothbort, 2015)
Alphabeticity is a measure of the relative location of a linguistic element in the alphabetical order and based on the findings above it provides a valid measure of fluency through which to corporate names.

2.3.4.4 Readability

The consideration of readability as a measure of fluency will provide a perspective that has not been addressed in the current studies of fluency in the context of corporate names and sentiment.

Fluency is a measure of the ease with which information is processed and understood. As such linguistic fluency of corporate names is most appropriately measured using the concept of readability, which is a statistical measure of the extent of mental development and cognitive effort required to process and understand language stimulus

The efficacy of measures of readability has been demonstrated in the business context in a variety of applications. Tan, Wang and Zhou(2013) found that positive sentiment expressed in company reports that have low readability is more likely to result in investor bias than positive sentiment in reports that have high readability. The Gunning Fog Score has been applied to automatically rank the potential usefulness, validity and appreciation of online reviews in an ecommerce environment (Korfiatis, García-Bariocanal, & Sánchez-Alonso, 2012). Most recently popular measures of readability such as the Flesch Kincaid Reading Ease and the Automated Readability Index have been employed in applications to automatically customise and personalise presentations in order to increase trust and persuasiveness (Khataei & Arya, 2015).

The traditional readability formulae (Flesch reading ease score, Automated readability index, Flesch-Kincaid grade level, Coleman-Liau index, Gunning fog index and SMOG index) are based on mathematical variants of formulas that consider factors such as the number of syllables, characters per word and complexity counts to score the readability. This rich source of fluency measures provides an additional lens through which to explore the effect of fluency of corporate name on the sentiment directed towards a company.

The preceding discussion (as is the case with sentiment) underscores the importance of testing a number of alternative measures in order to fully understand the potential impact of fluency.

This section discusses the overall the themes identified in the analysis of literature that is pertinent to researching the effect of fluency of corporate name on social media sentiment toward a company.

The study of sentiment in a social media setting is important because it provides an opportunity to compare sentiment effects in data produced in an environment consisting of millions of “independent” agents and data generated from consulting a panel of experts or opinion makers.

3 Research Questions

3.1 Introduction

The literature reviewed shows that the existence of cognitive bias in the sentiments expressed on stocks is present and has been studied for a long while. However the fact that the sentiments expressed (whether on social media or by other analysts) are predictive of future performance of stocks makes it clear that on the whole stock sentiments are driven by fundamental performance of the companies on which opinions are expressed – although cognitive bias (such as may be created by factors such as fluency of corporate name) will mediate this causatory relationship.

The basic research question, therefore, is that fluency of corporate name is mediator of the relationship between fundamental performance and stock sentiment data.

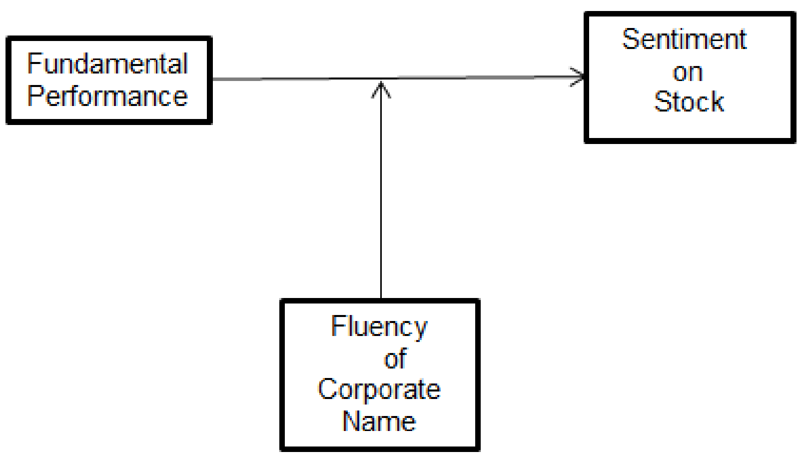


Figure 1: The Research Question

The research question is further elaborated in the research questions below.

3.2 Research Question 1

As a preliminary to understanding the mediating effect of fluency on stock sentiments, it is of interest to establish whether the distribution of sentiments expressed differs between social media and investment analysts.

Question:
The diversity of opinion obtained from social media differs from that obtained from investment experts

3.3 Research Question 2

A key insight from the literature is that crowd sourced intelligence is less effective in the presence of imitation and centrality. The extent of these variables differs between the social media sentiment sources.

Question:
The concentration of opinion and investment advice differs between social media sources

3.4 Research Question 3

The fluency of corporate name is a mediator of the effect of fundamental performance on investment analyst sentiment

3.4.1 Research Question 3.1

The fluency of corporate name is a mediator of the effect of fundamental performance on investment analyst sentiment, and the effect differs for the different measures of fluency

3.5 Research Question 4

The fluency of corporate name is a mediator of the effect of fundamental performance on social media sentiment

3.5.1 Research Question 4.1

The fluency of corporate name is a mediator of the effect of fundamental performance on social media sentiment, and the effect differs for different sources of social media sentiment data

4 Research Methodology

4.1 Introduction

This research investigated the extent to which corporate name fluency mediates the relationship between company fundamental performance and social media sentiment towards the company.

The contribution it aims to make is to quantify the extent of cognitive bias introduced into stock sentiments expressed on both social media and by investment analysts as a result of corporate name fluency.

A comparative analysis was done by analysing historical stock sentiment data and calculating the fluency of corporate name of each of the stocks considered in order to correlate the sentiment data to corporate name fluency.

Separate data sources were (necessarily) used to source the investor analyst sentiment data and the social media sentiment data; however the same measures of fluency were tested on the sentiment data obtained.

4.2 Choice of methodology

Due to the volume of data considered and the quantitative nature of the variables being investigated, a purely quantitative research methodology was used. Regression modelling techniques were used to draw statistically significant inferences on the extent to which corporate name fluency mediates the relationship between fundamental corporate performance and sentiment expressed towards that corporate (see Section 4.5 below)

4.3 Population and Sampling

4.3.1 Unit of analysis

The unit of analysis for the research will be the individual companies in the stock sentiment database.

4.3.2 Population

The population of relevance consists of all companies with equities listed on the NYSE and NASDAQ during the 12 month period from January to December 2014.

4.3.3 Sampling size

Two separate data samples were used in the research, one for investment analyst sentiment data and another for social media sentiment data. An additional data source was used to collect information on underlying fundamental performance of the equities researched.

The investment analyst sentiment data was sourced from Instititutional Brokers Estimates System (IBES) database that is maintained by Thomson Reuters (see 4.3.3.1 below).

The social media sentiment data was obtained Psych Signal a leading vendor of social media derived financial sentiments that maintains a database of daily Twitter derived sentiment for "more than 10,000 financial market products, including stocks, ETFs, futures, currencies and even bitcoin. Updated daily, this is the largest database of its kind in the world, with detailed history going back to 2009"

4.3.3.1 IBES data sample

The IBES data sample consists of the following fields:

- The number of analyst recommendations
- The number of buy recommendations
- The number of sell recommendations
- The number of hold recommendations (if you do not have it don't buy but if already invested do not sell)

4.3.3.2 Psych Signal Data

The daily financial sentiment data provided by PsychSignal includes 6 years of history of the following columns for each stock:

Table 1: Psych Signal Sentiment measures

Column	Description
Date	This is the date of the analyzed data.
Bullish Intensity	The algorithms score each message's language for the strength of bullishness present on a 0-4 scale. 0 indicates no bullish sentiment measured, 4 indicates strongest bullish sentiment measured. 4 is rare.
Bearish Intensity	The algorithms score each message for the strength of bearishness present in the message on a 0-4 scale. 0 indicates no bearish sentiment measured, 4 indicates strongest bearish sentiment measured. 4 is rare.
Bull - Bear	This indicator simply subtracts bearish intensity from bullish intensity to provide an immediate net score.
Bullish Messages	This indicator is the total count of bullish sentiment messages scored by the algorithm.
Bearish Messages	This indicator is the total count of bearish sentiment messages scored by the algorithm.
Total Messages	This indicator is the number of messages coming through our source data feeds and attributable to a symbol regardless of whether our sentiment engine can score them for bullish or bearish intensity.

Note. Retrieved from https://www.quandl.com/data/PS1/documentation/documentation.

The financial sentiments data are further classified along the following sub-categories:

1. only StockTwits
2. only Twitter without retweets
3. only Twitter with retweets
4. Twitter without retweets aggregated with StockTwits
5. Twitter with retweets aggregated with StockTwits.

The StockTwits classification is of particular interest as it isolates twitter streams generated by the StockTwits platform, which is a microblogging community that is dedicated to stock markets and comprises of c. 150,000 members that exchange and alert on market info (Oliveira, Cortez, & Areal, 2013). This feed of data is possibly less decentralised and independent than the rest of the Twitter feed and it is therefore useful to compare the distribution of sentiment and influence between this and indeed amongst all the other classifications. The "with or without retweets" classification for example allows for a test of the extent of imitation and centralisation of opinion in the data.

4.3.4 Sampling method / technique

The entire population of financial sentiments expressed on NYSE and NASDAQ listed equities during the 12-month period between January and December 2014 was included in the data set that was analysed.

4.4 Research variables

This section includes a description of the variables used to measure fluency, sentiment and fundamental financial performance.

4.4.1 Measures of fluency

Eight measures of fluency were calculated for each equity in the data sample as shown below.

4.4.1.1 The Corpus Percentile

The corpus percentile was calculated from downloading the entire Google corpus of language and applying a percentile rank in terms of frequency of usage to each word that appears in that trillion token dataset (see Section **Error! Reference source not found.**). Each company name was assigned a corpus percentile by calculating a corpus percentile for each word in the name(in the case of multi word names) with the final ranking being the percentile of the highest ranked element (higher ranking indicating reduced frequency of use).

4.4.1.2 Alphabeticity

The alphabeticity was calculated by ranking each of the companies in the population (NYSE and NASDAQ) listed in alphabetical order and assigning equity a percentile (0 to 100) based on its position in that rank ordering, higher ranking indicating distance away from 'A'.

A number of statistical readability indices were calculated for each equity based on parsing each name through each of the readability formulae below. The task was completed for all the company names in the study using an online subscription based readability calculator that returns a score on each dimension below for submitted text (**https://readability-score.com/**). The readability formulae are calculated as follows:

The Flesch Kincaid Reading Ease :

FRES = 206.835 - (1.015 * wordCount) / sentenceCount - (84.6 * syllable Count) / wordCount

The Flesch Kincaid Grade Level

FKGL = (0.39 * wordCount) / sentenceCount + (11.8 * syllableCount) / wordCount - 15.59

Gunning Fog Score

FOG = 0.4 * ((double)wordCount / sentenceCount + (100.0 * complexCount) / wordCount);

Coleman Liau Index

CL = (5.89 * letterNumberCount) / wordCount - (30.0 * sentenceCount) / wordCount - 15.8

SMOG Index

SMOG = SQUARE ROOT(complexCount * 30.0 / sentenceCount) + 3.0

Automated Readability Index

ARI = (4.71 * letterNumberCount) / wordCount + (0.5 * wordCount) / sentenceCount -21.43

4.4.2 Measures of fundamental performance

The following variables (as at Q4 2014) were used to test for the relationship of sentiment to underlying fundamental performance:

- The 3 year average earnings per share percent
- The 3 year average net income percent
- The asset turnover ratio
- The book value per share
- The return on invested capital percent
- The total revenue
- The total current liabilities percentage
- The total stockholder equity percent
- The year over year revenue percent
- The net margin percent
- The financial leverage as a percentage of profitability

A regression model against sentiment was fitted and the variables were collapsed into a single measure by replacing them with fitted value based on multiply each value with its estimated parameter and obtaining the (fundamental) fitted value.

4.4.3 Data transformations

In order to facilitate the comparison of data measured in different units all of the research variables were transformed into percentile ranks ranging 0 to 100. Thus for example instead of trying to predict the exact number of bullish sentiments expressed towards a stock, the stocks were percentile ranked in terms of the total bullish sentiments they received and this percentile ranking was the dependent variable predicted by the model. Similarly the predictor effects in the model were converted into percentile rankings i.e. in terms of underlying fundamentals and in terms of position on each of the 8 fluency measures.

4.5 Data analysis

The SAS software system (Statistical Analysis System Institute, 1999) was used to perform the data analysis and generate distribution analysis, regression models and diagnostics required to test the research questions.
This section describes the statistical methodology used to test each of the research questions.

4.5.1 Distribution analysis

Research questions 1 and 2 ask whether the variability and independence of opinions sourced from investment analysts differs from that obtained from social media. The question looks for differences in the diversity of sentiment expressed.

In order to test the diversity of opinion the sentiment datasets were filtered to look exclusively at equities for which at least one positive sentiment had been expressed in the last quarter of 2014 and then analysing what the overall distribution of sentiment towards stocks was within the data.

4.5.1.1 Histogram analysis

The first element of testing distribution of sentiment data was to calculate for each stock the proportion of overall sentiment that was positive for stocks that had at least one positive opinion expressed during the period and analysing whether this distribution was bell shaped , which would indicate a spread of opinions or if the distribution was highly skewed to one or other extreme which would indicate levels of uniformity or lack of spread. The variability and spread of opinion were quantified using the skewness and kurtosis statistics described in Section 4.5.1.2 below.

4.5.1.2 Skewness and kurtosis

The skewness of a data sample is a statistical measure of the extent and direction of asymmetry in the data.

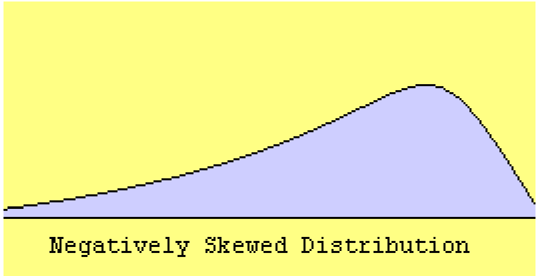


Figure 2: Negatively skewed distribution

(Note: Retrieved from http://www.ats.ucla.edu/stat/sas/output/univ.htm)

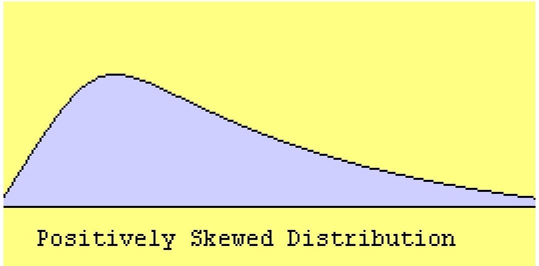


Figure 3: Positively skewed distribution

(Note: Retrieved from http://www.ats.ucla.edu/stat/sas/output/univ.htm)

Distributions with a normal or bell shaped distribution have a skewness of 0, positively skewed distributions have a skewness greater than 0 and negatively skewed distributions have skewness less than 0. The extent of asymmetry can gaged and compared based on the magnitude of the skewness value.

The kurtosis measures the weight of observations in the tails of a distribution. Normally distributed data will have kurtosis close to zero whilst non-normality is measured by the extent of difference (negative or positive values) from zero.

4.5.2 Moderated regression analysis

Moderated regression analysis measures how the relationship between two variables measured in a regression model, varies as a result of a third variable called an interaction or moderation variable.

4.5.2.1 Main and Interaction effect regression models

The inputs into the moderated regression model consist of three elements (Statistical Analysis System Institute, 1999):

- The independent variable (sentiment in this case)
- The primary dependent variable (fundamental financial performance)
- The moderator variable (fluency of corporate name)

4.5.2.2 Model specification

The regression model proceeds by testing and quantifying 3 relationships in succession:

- Independent variable = Primary dependent
- Independent variable = Primary dependent + Moderator Variable
- Independent variable = Primary dependent + Moderator Variable + Interaction

The interaction term is simply the product of the primary dependent and moderator variables.

The primary and moderator variables are referred to as main effects and the interaction variable as the interaction effect.

The existence of a statistically significant moderation relationship is tested by fitting a main and interaction effects regression model. If the coefficient of the interaction term is statistically significant (p-value less than 0.05) and the overall predictive "power" of the model increases when the interaction term is included, this proves the existence of a moderating effect. The overall predictive power of the model is measured by the R^2 value with a higher value indicating better model performance.

Finally the F-Ratio test provides further evaluation of the statistical significance of the interaction relationship. A statistically significant F-Ratio in the interaction effects regression model confirms the existence of an interaction effect.

4.5.2.3 Coefficient and Effect Plots

The nature and direction interaction effect can be examined visually using effect and coefficient plots. The effect plots show the slope of the relationship between the independent and primary dependent variables at a number of ranges of the interaction variable which provides clear visual evidence of the nature and direction of the relationship. An example of an effect plot displaying the moderating effect on the relationship between mathematical ability and anti-social behaviour in children of hyperactivity at 3 different levels (low, mean and high).

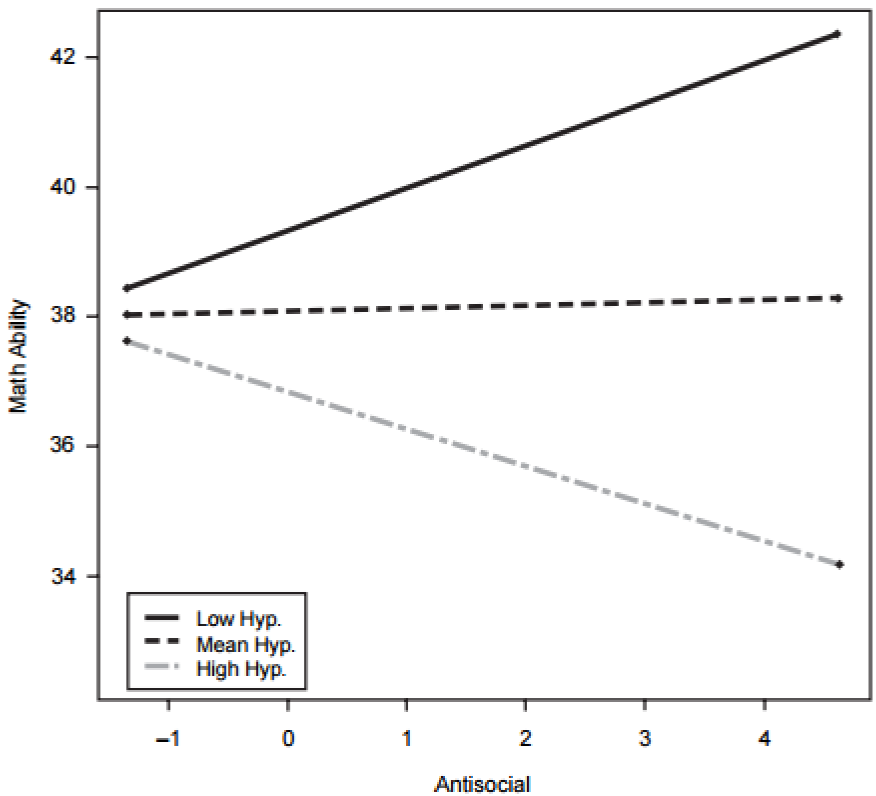


Figure 4 Effects plot

Coefficient plots graphically display the distribution of estimated coefficients of the primary dependent variable at different levels of the moderator variable.

4.5.2.4 Model diagnostics (reliability and validity)

The regression model analysis produces a number of diagnostic reports that are examined to ensure that the inferences drawn from the model outputs are valid and reliable (Statistical Analysis System Institute, 1999). The examination of these model diagnostics is critical as it determines the applicability or potential spuriousness of the results obtained from the statistical analysis.

The following model diagnostic outputs are examined:

The residual and externally studentized residuals (RSTUDENT) versus the predicted values and versus leverage

The residual is the model error (actual value – predicted value) the studentized residual is the residual divided by its standard error. The values in these plots should ideally be randomly scattered and not display clear patterns.

The leverage plot shows the extent of outlying (and influential) observations in the data as indicated by the number of observations at an extreme distance outside of the central left box in the far right panel below.

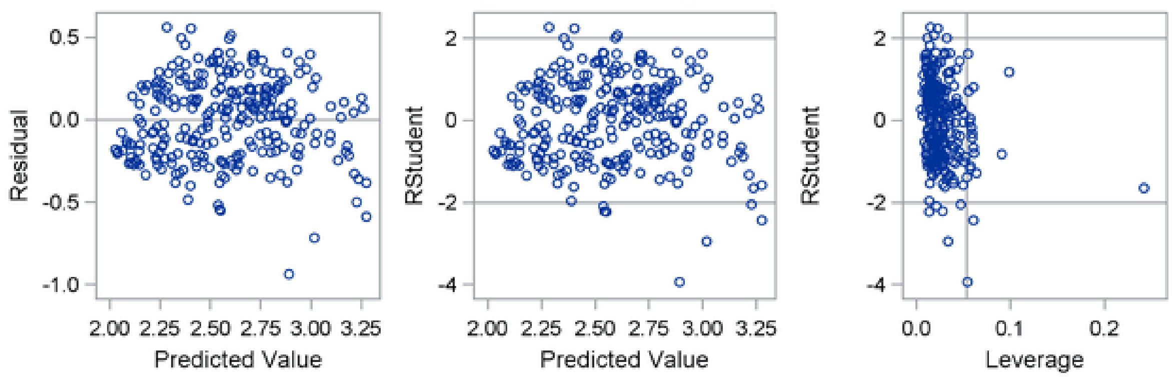


Figure 5: Residual plots ((Statistical Analysis System Institute, 1999))

The residual versus quantile plot, the observed versus predicted plot should lie as close as possible to the 45 degree line in the central panel. Cook's D is a measure of the effect on the model of deleting a single observation, and values that breach the threshold for Cook's D shown in the far right panel below are generally influential observations with significant influence on the model.

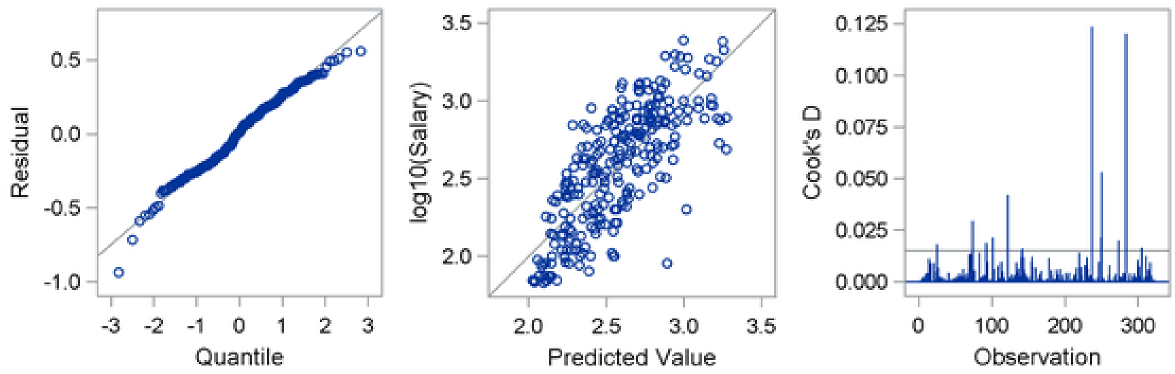


Figure 6: Residual v Quantile, Observed v Predicted and Cook's D plots

The plots below show the histogram of prediction errors on the left hand side and a comparison of the centered mean prediction value (fit-mean is the model prediction less the mean prediction) to the standardised residual (to also have zero mean) model prediction error.

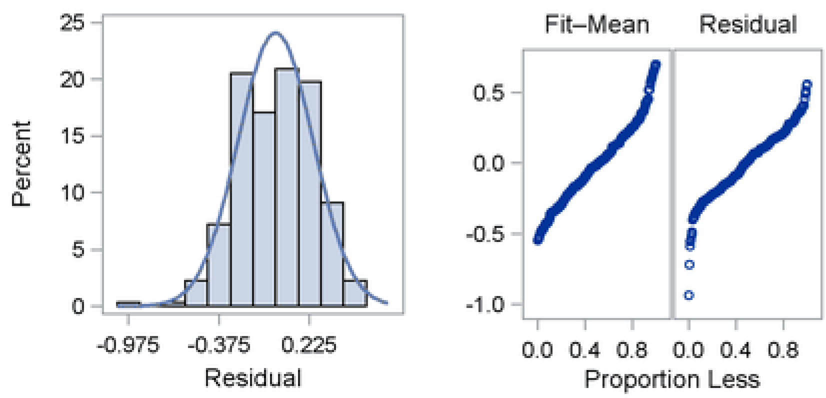


Figure 7: Normal test and Fit Diagnostics

Ideally the residuals (model errors in the left panel) should be normally distributed. The double panel on the right compares the spread of the fitted values to the spread of the errors. Extreme values in the left panel indicate influential observations that have leverage in terms of swinging the overall model estimated by the regression, outliers in the left panel indicate outliers for which the actual value is very different from the predicted value.

4.5.3 Methodology limitations

The methodology applied is quantitative in nature and applies techniques that are effective to empirically identify the extent of moderation in the relationship between fundamental performance and sentiment because of fluency of corporate name.

The model however is reliant on the accuracy of data supplied and the completeness of predictors included. The present model only includes fundamental performance and fluency of corporate name as predictors of sentiment and therefore to the extent that other significant predictors of sentiment have been omitted from the model, it will be limited and errors will be systematically introduced.

4.6 Conclusion

The methodology applied to test the research questions and explanation of the model variables and sample selection were described in this chapter, including the possible limitations of the research.

The next chapter presents the results of the research.

5 Presentation of Research Results

5.1 Introduction

5.2 A description of the sample obtained

The final sample analysed was comprised as follows

Social Media Sentiment Data

Sector	
BASIC MATERIALS	2427.
CONGLOMERATES	75.
CONSUMER GOODS	1721
FINANCIAL	4472
HEALTH CARE	2101
INDUSTRIAL GOODS	1484
SERVICES	3741
TECHNOLOGY	3458
UTILITIES	565
Total	20044

Investment analyst data

Sector	
BASIC MATERIALS	315.00
CONGLOMERATES	10.00
CONSUMER GOODS	245.00
FINANCIAL	601.00
HEALTH CARE	313.00
INDUSTRIAL GOODS	217.00
SERVICES	569.00
TECHNOLOGY	506.00
UTILITIES	98.00
Total	2874.00

5.3 Test for differences in performance between fluency measures and channels of social media data

The objective of research questions 3 and 4 was not only to test whether fluency mediates the relationship between fundamentals and sentiment , but also whether this effect differed for different measures of fluency and different sources of stock sentiment data (in the case of social media sentiment)

The questions were stated as follows:

- The fluency of corporate name is a mediator of the effect of fundamental performance on investment analyst sentiment, and the effect differs for the different measures of fluency (Question 3.1)

- The fluency of corporate name is a mediator of the effect of fundamental performance on social media sentiment, and the effect differs for different sources of social media sentiment data(Question 4.1)

In order to test for the differences in effect across different measures of fluency and different sources of social media data, a regression model to predict the level of sentiment including fundamental performance and each one of the measures of fluency and for each classification of social media data (in the case of the social media data set) was performed and the r-squared values obtained compared to identify any differences in model performance. The r-squared ranges from 0 to 1 and approximates the proportion of variance in the predicted variable that is explained by the model.

The graph below shows the r-squared values obtained from the social media data when individually fit in different combinations with the fundamental performance.

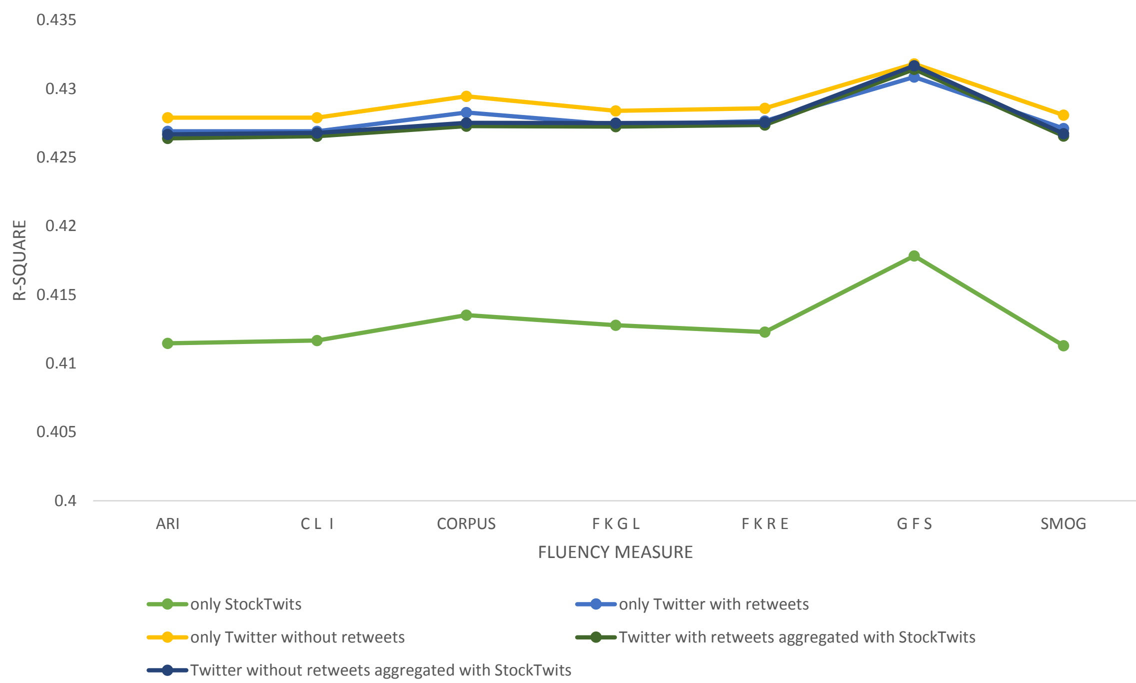


Figure 8 R-Squared comparison for social media fluency and fundamental performance models of sentiment

The graph above delivers a fascinating insight that the StockTwits data generated by the online community of investment enthusiasts has the least predictive power on sentiment and that the Gunning Fog Score appears to have superior predictive impact on sentiment relative to the other measures.

Figure 9 R-Squared comparison for fluency and fundamental performance models of sentiment (Investment Analyst Data)

The value of r-squared between the different fluency measures appears to be narrow for the investment analyst data, however it is interesting to note that the trend of the Gunning Fog Score performance appears to be reversed in the analyst data.

5.4 Distribution analysis

The distribution analysis asks a simple question – if an equity received at least one positive recommendation in a specific medium, what proportion of the messages expressed on it in that medium were positive. The range of answers to that question in each medium is plotted below. The distribution analysis was performed to establish what proportion of all sentiment expressed was positive for the population of stocks that received positive recommendations in the last quarter of 2014.

A comparison of three sentiment data streams was made:

- Investment broker data
- StockTwits social media tweets
- The overall twitter feed – Twitter with Retweets Aggregated with StockTwits

The graph below shows the distribution of proportion of sentiment that is positive for positive recommendations made on Twitter. The mean value is that 75% of all messages will be positive for a company that has any positive recommendations made. There is a bimodal distribution in the overall trend with a significant sub-group (about 7,5%) receiving exclusively and a tail of observations (possible gold nuggets) that receive a tiny proportion of positive recommendations.

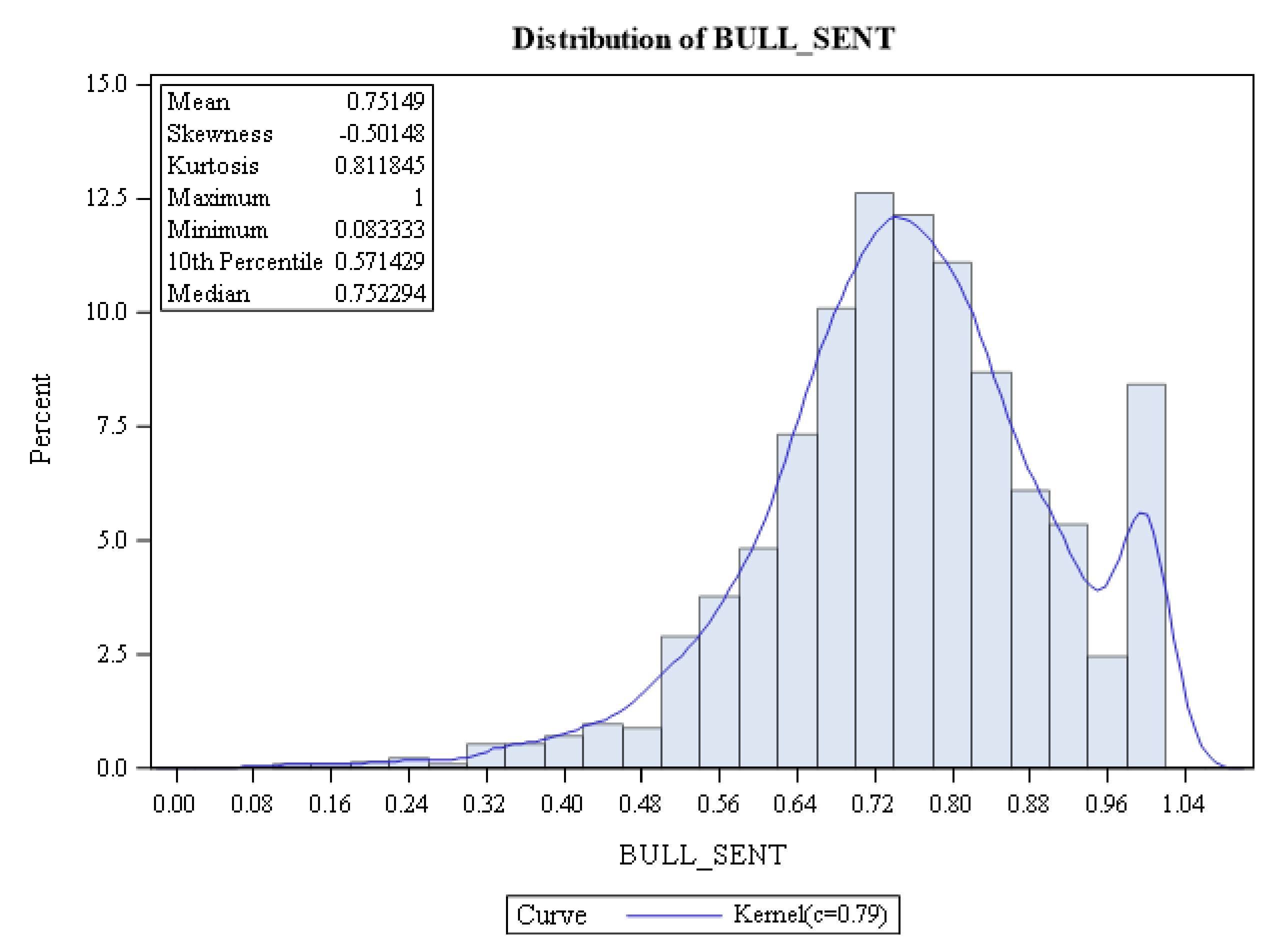


Figure 10: Proportion of positive sentiment StockTwits

This is compared to the aggregate Twitter feed below.

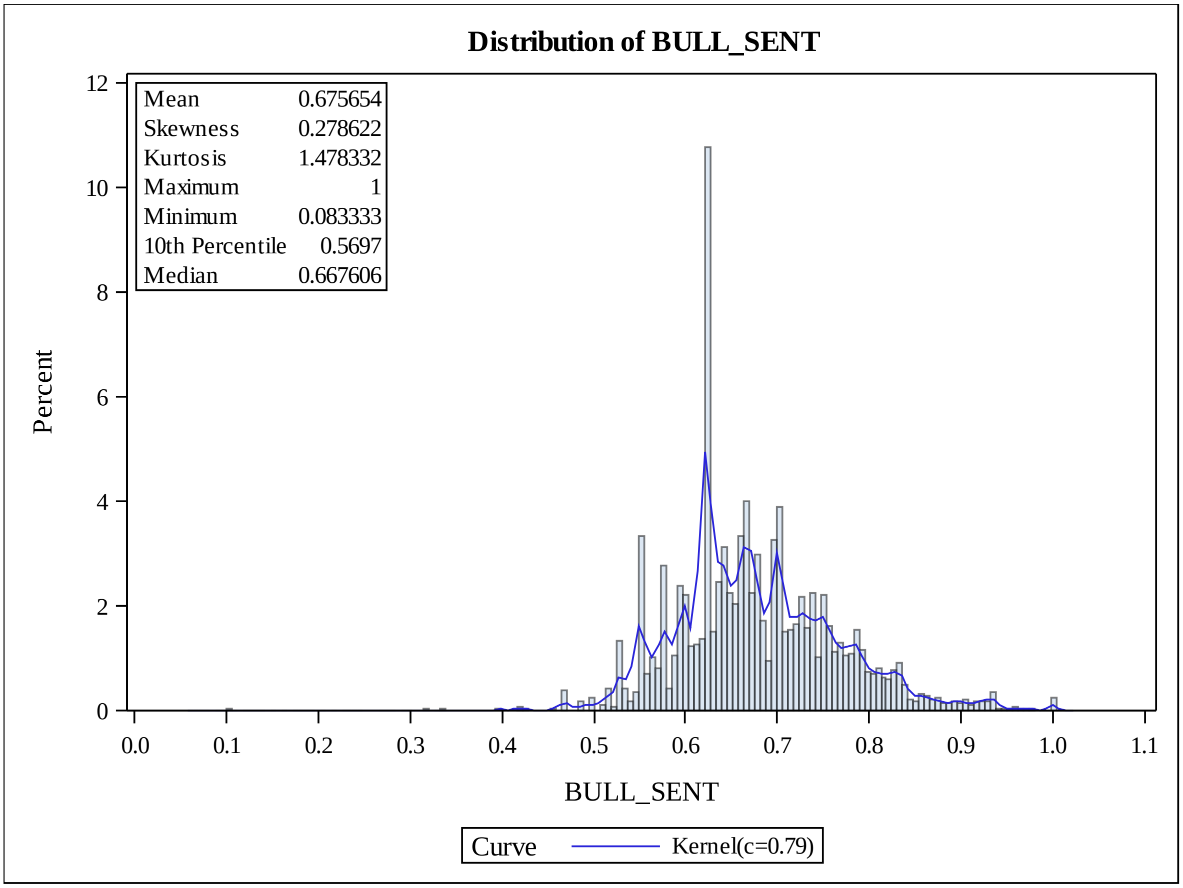


Figure 11: Proportion of positive sentiment Twitter with Retweets Aggregated with Stock Twits

The distribution of recommendations in the overall Twitter feed for stock that receive at least one positive recommendation is much more normal shaped and therefore random or varied.

The final graph (below) shows the distribution of proportion of positive sentiment expressed for equities that have received at least one positive recommendation in the investment broker data set.

In this case it is striking that such equities will receive almost exclusively positive sentiment and that there is significantly less variance of views.

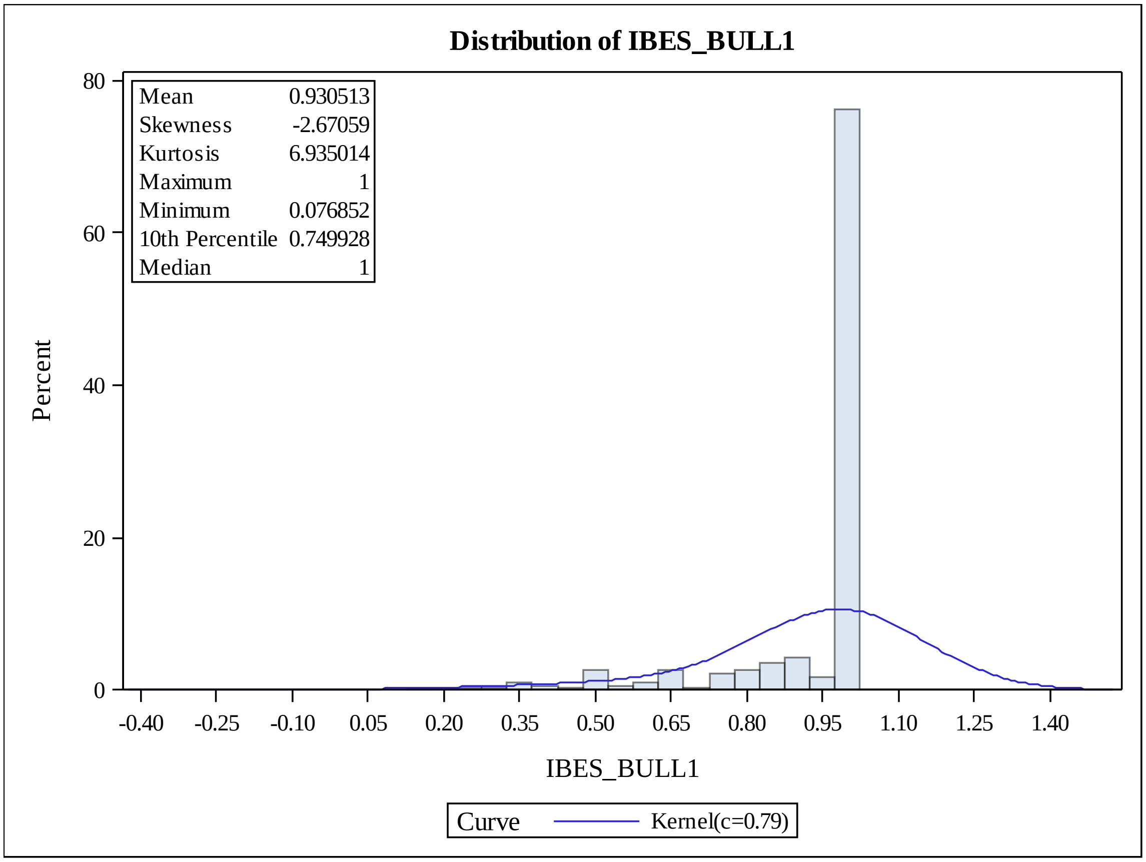


Figure 12: Proportion of positive sentiment , investmeent broker and analyst data

5.5 Results on reliability and validity of the data

Detailed diagnostics were produced for each regression model. These are reviewed below as a prelude to reviewing the findings of the moderated regression analysis. The first section below considers the diagnostics for the models estimated for investment analyst data.

5.5.1 Diagnostic plots Research Question 3 – Investment Analyst Data

The panel below shows the model diagnostics for the main effects model for corpus percentile and financial fundamentals (as the predictors) and sentiment as the predicted outcome for investment analyst data. The first two panels exhibit a clear downward slope in the residuals suggesting that the model is better at predicting at the extreme end of the distribution (those that will receive in the top percentiles of positive sentiment and less so for that get less. The top right hand and middle right hand panel also point to the existence of a number of outliers (see Section 4.5.2.4).

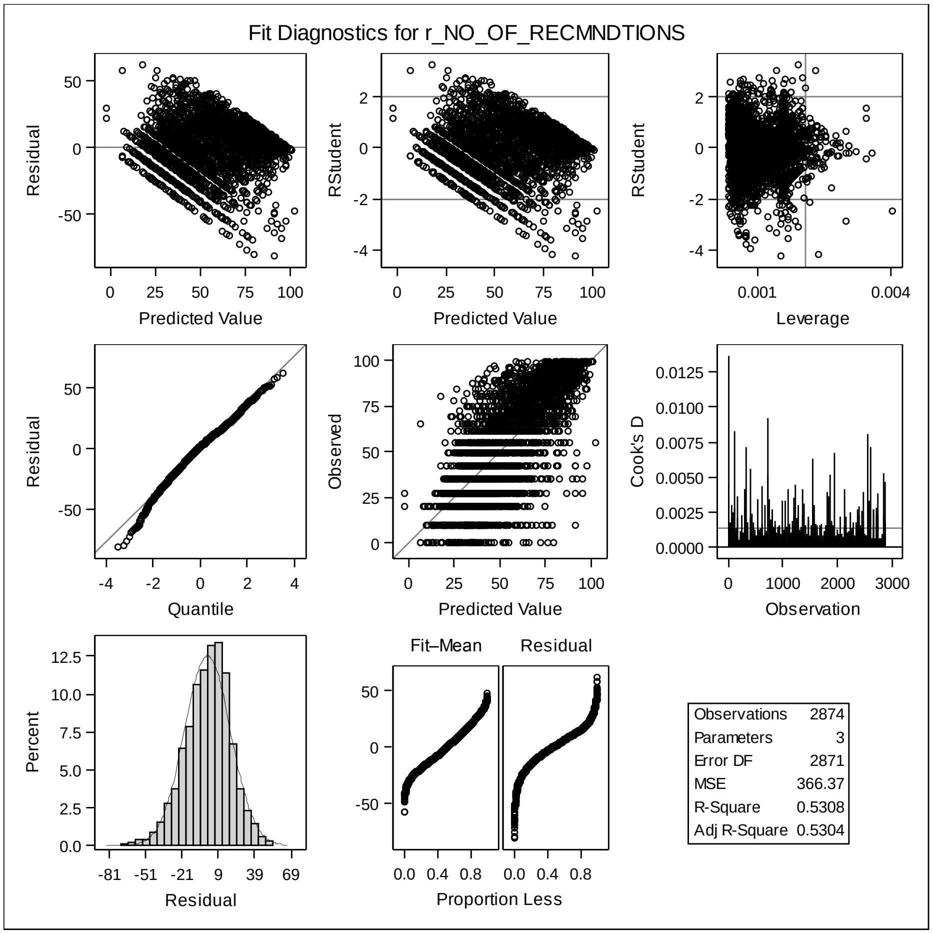


Figure 13 - Diagnostic Plots - Corpus Percentile (IAD)

The plot shows that outliers are more prevalent than individual data points that have influence on the estimated equation. The overall model fit is decent however as the residuals have a normal distribution in general and as the studentized residuals (see the top middle panel) generally lie within (-2,2) recommended cut-offs . The diagnostic trend is similar in the main effects models for the other fluency measure which are shown in Appendix below.

5.5.2 Diagnostic plots Research Question 3 – Social Media Data

Diagnostic analysis was also performed for the main and interaction effect models fitted to the social media sentiment data. Similar to the trend observed in the investment analyst model - it was apparent that instead of the model errors being random without an evident pattern (as would be the case if they all fell in a random scatter along the horizontal axis in the top left hand corner of the panel below), the model errors reduced as the percentile of predicted value increases.. The diagnostics show that the model is better at predicting companies that will be in the top percentiles of sentiment than it is at predicting those that will not. Overall whilst the diagnostics exhibited the existence of a significant numbers of outliers and influential observations (top right hand and middle right hand panels) the overall requirement of normally distributed errors (bottom left hand panel) was satisfied allowing us to examine moderated regression analysis models (bearing in mind the diagnostics identified here. The plot below shows the diagnostic plot for corpus percentile as a fluency measure. A similar diagnostic trend was observed for the other measures of fluency which are shown in Appendix below.

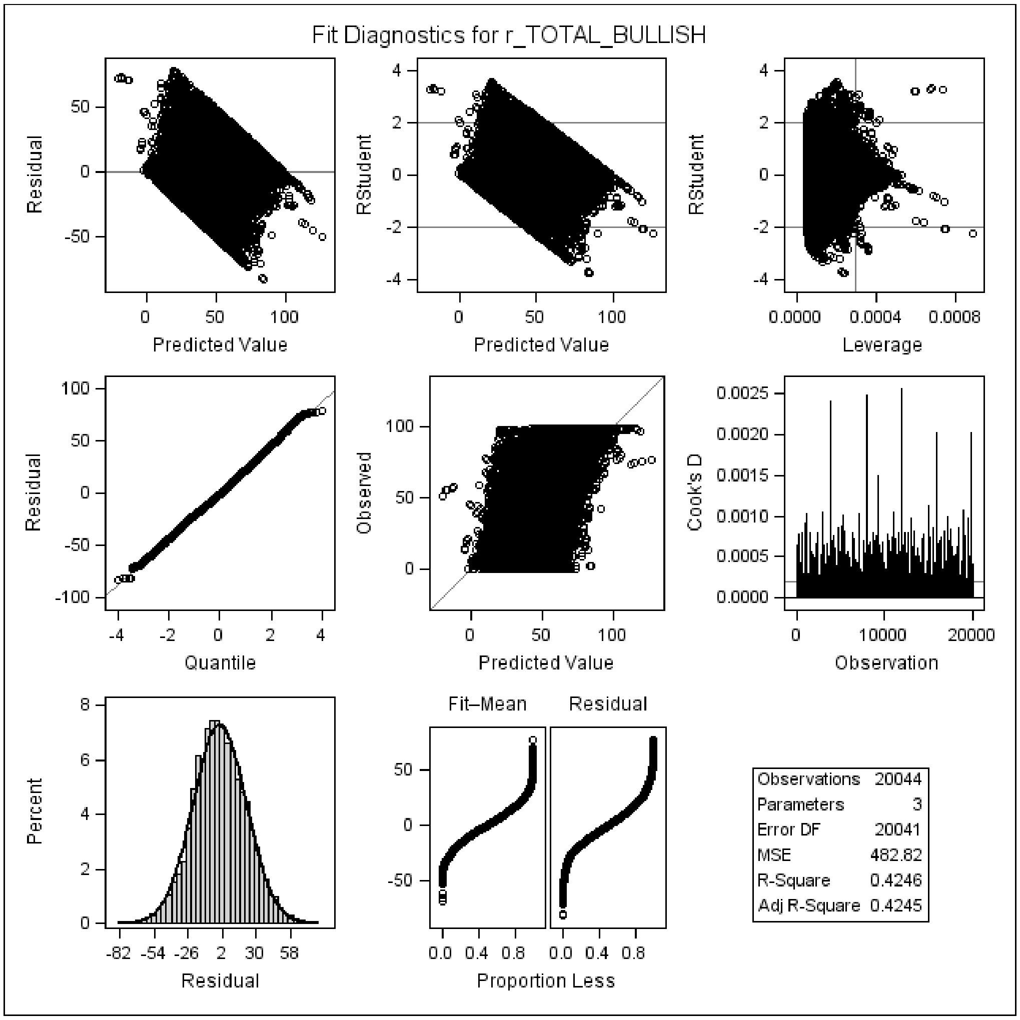


Figure 14 - Diagnostic Plots - Corpus Percentile SMD

5.6 Moderated Regression Analysis (Investment Analyst Data)

5.6.1 Main effect model (Investment Analyst Data)

The main effects model tests for a statistically significant relationship between the main effects (fundamental performance and corpus percentile below) and sentiment. The overall model is significant for both main effect (see p-value Pr > |t| below less 0.05 for each) with an over all model r-squared of 0.53.

Table 2 - Main effects model– Corpus Percentile (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2874

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1189738	594869	1623.70	<.0001
Error	2871	1051840	366.36718		
Corrected Total	2873	2241578			

Root MSE	19.14072	R-Square	0.5308
Dependent Mean	55.04523	Adj R-Sq	0.5304
Coeff Var	34.77271		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	-0.99014	1.14805	-0.86	0.3885	0	0	-3.24123	1.26096
fundamental		1	1.00707	0.01793	56.18	<.0001	0.73325	1.04236	0.97192	1.04222
corpus_pctl		1	0.01904	0.00974	1.96	0.0506	0.02552	1.04236	-0.00005257	0.03814

Whilst the overall model is significant, the main effect model results indicate that alphabeticity is not a statistically significant predictor of investment analyst recommendations.

Table 3 - Main effects model – Alphabeticity (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2874

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	1189134	594567	1621.94	<.0001
Error	2871	1052444	366.57742		
Corrected Total	2873	2241578			

Root MSE	19.14621	*R-Square*	0.5305
Dependent Mean	55.04523	*Adj R-Sq*	0.5302
Coeff Var	34.78269		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	0.86821	1.18695	0.73	0.4646	0	0	-1.45914	3.19557
fundamental		1	1.00009	0.01756	56.94	<.0001	0.72817	1.00001	0.96565	1.03453
r_ALPHABETICITY	Rank for Variable ALPHABETICITY	1	-0.01799	0.01220	-1.47	0.1404	-0.01886	1.00001	-0.04190	0.00593

The Flesch Kincaid Reading Ease is also not a statistically significant predictor of the level of sentiment.

Table 4 - Main effects model – Flesch Kincaid Reading Ease (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	1186055	593027	1617.57	<.0001
Error	2867	1051091	366.61712		
Corrected Total	2869	2237146			

Root MSE	19.14725	*R-Square*	0.5302
Dependent Mean	55.08118	*Adj R-Sq*	0.5298
Coeff Var	34.76187		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	0.38176	1.13841	0.34	0.7374	0	0
fundamental		1	1.00081	0.01767	56.64	<.0001	0.72901	1.01106
r_Flesch_K_R_E	Rank for Variable Flesch_Kincaid_Reading_Ease	1	-0.00861	0.01233	-0.70	0.4852	-0.00898	1.01106

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-1.85042	2.61394
fundamental		1	0.96616	1.03546
r_Flesch_K_R_E	Rank for Variable Flesch_Kincaid_Reading_Ease	1	-0.03279	0.01557

Similarly the Flesch Kincaid Grade Level is not a statistically significant main effect.

Table 5 - Main effects model – Flesch Kincaid Grade Level (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	1186094	593047	1617.68	<.0001
Error	2867	1051052	366.60342		
Corrected Total	2869	2237146			

Root MSE	19.14689	*R-Square*	0.5302
Dependent Mean	55.08118	*Adj R-Sq*	0.5299
Coeff Var	34.76122		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	-0.46902	1.22872	-0.38	0.7027	0	0
fundamental		1	1.00096	0.01767	56.64	<.0001	0.72912	1.01137
r_Flesch_K_G_L	Rank for Variable Flesch_Kincaid_Grade_Level	1	0.00941	0.01221	0.77	0.4408	0.00993	1.01137

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-2.87827	1.94024
fundamental		1	0.96631	1.03562
r_Flesch_K_G_L	Rank for Variable Flesch_Kincaid_Grade_Level	1	-0.01453	0.03336

The Gunning Fog Score is not a statistically significant main effect.

Table 6 -Main effects model – Gunning Fog Score (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	1185920	592960	1617.18	<.0001
Error	2867	1051226	366.66411		
Corrected Total	2869	2237146			

Root MSE	19.14848	*R-Square*	0.5301
Dependent Mean	55.08118	*Adj R-Sq*	0.5298
Coeff Var	34.76410		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	-0.10428	1.11894	-0.09	0.9258	0	0	-2.29828	2.08973
fundamental		1	0.99971	0.01758	56.85	<.0001	0.72821	1.00102	0.96523	1.03419
r_Gunning_F_S	Rank for Variable Gunning_Fog_Score	1	0.00415	0.01201	0.35	0.7293	0.00443	1.00102	-0.01939	0.02769

The test also shows that the Coleman Liau Index is not a statistically significant main effect.

Table 7: Main effect model CLI – (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	1186772	593386	1619.65	<.0001
Error	2867	1050374	366.36699		
Corrected Total	2869	2237146			

Root MSE	19.14072	*R-Square*	0.5305
Dependent Mean	55.08118	*Adj R-Sq*	0.5302
Coeff Var	34.75001		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	-1.02626	1.23829	-0.83	0.4073	0	0	-3.45430	1.40177
fundamental		1	1.00210	0.01765	56.79	<.0001	0.72995	1.00882	0.96750	1.03670
R_Coleman_L_I	Rank for Variable Coleman_Liau_Index	1	0.01932	0.01236	1.56	0.1180	0.02010	1.00882	-0.00491	0.04354

The SMOG Index is not a statistically significant main effect

Table 8 - Main effects model – SMOG Index (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	1186042	593021	1617.53	<.0001
Error	2867	1051104	366.62161		
Corrected Total	2869	2237146			

Root MSE	19.14737	*R-Square*	0.5302
Dependent Mean	55.08118	*Adj R-Sq*	0.5298
Coeff Var	34.76208		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	0.35058	1.12684	0.31	0.7557	0	0	-1.85892	2.56008
fundamental		1	0.99859	0.01763	56.65	<.0001	0.72740	1.00610	0.96403	1.03316
r_SMOG_Index	Rank for Variable SMOG_Index	1	-0.00829	0.01233	-0.67	0.5014	-0.00863	1.00610	-0.03246	0.01588

Similarly the Automated Readability Index is not a significant main effect.

Table 9 - Main effects model – Automated Readability Index (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1186768	593384	1619.64	<.0001
Error	2867	1050378	366.36849		
Corrected Total	2869	2237146			

Root MSE	19.14075	R-Square	0.5305
Dependent Mean	55.08118	Adj R-Sq	0.5302
Coeff Var	34.75008		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	1	-1.02842	1.23998	-0.83	0.4070	0	0
fundamental		1	1.00217	0.01765	56.78	<.0001	0.73000	1.00934
r_Auto_R_I	Rank for Variable Automated_Readability_Index	1	0.01921	0.01232	1.56	0.1189	0.02005	1.00934

Parameter Estimates

Variable	Label	DF	95% Confidence Limits	
Intercept	Intercept	1	-3.45976	1.40292
fundamental		1	0.96756	1.03677
r_Auto_R_I	Rank for Variable Automated_Readability_Index	1	-0.00494	0.04336

The analysis shows that out of all 8 fluency measures tested as possible main effects in a join model to predict sentiment with financial fundamentals, only the corpus percentile was a statistically significant main effect. The next section considers the interaction effects in the models.

5.6.2 Interaction Effects Model (Investment Analyst Data)

The analysis indicates that the interaction term between fundamental performance and corpus percentile is statistically significant and that overall model fit (r-squared) improves with this variable included.

Table 10 - Interaction Effects Model – Corpus Percentile (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2874

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	1192164	397388	1086.80	<.0001
Error	2870	1049414	365.64957		
Corrected Total	2873	2241578			

Root MSE	19.12197	*R-Square*	0.5318
Dependent Mean	55.04523	*Adj R-Sq*	0.5314
Coeff Var	34.73864		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	-2.93408	1.37296	-2.14	0.0327	0	0
fundamental		1	1.04265	0.02262	46.10	<.0001	0.75916	1.66254
corpus_pctl		1	0.08528	0.02749	3.10	0.0019	0.11430	8.32419
fundamental_corpus_pctl		1	-0.00133	0.00051600	-2.58	0.0101	-0.09331	8.04517

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-5.62617	-0.24198
fundamental		1	0.99830	1.08700
corpus_pctl		1	0.03137	0.13919
fundamental_corpus_pctl		1	-0.00234	-0.00031732

The interaction term for alphabeticity is not statistically significant at the 5% level however it could be accepted at the 10% level which is used in some applications

Table 11 - Interaction effects model – Alphabeticity (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2874

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	1189685	396562	1081.98	<.0001
Error	2870	1051893	366.51341		
Corrected Total	2873	2241578			

Root MSE	19.14454	*R-Square*	0.5307
Dependent Mean	55.04523	*Adj R-Sq*	0.5302
Coeff Var	34.77965		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	2.80003	1.97338	1.42	0.1560	0	0
fundamental		1	0.96521	0.03345	28.86	<.0001	0.70277	3.62753
r_ALPHABETICITY	Rank for Variable ALPHABETICITY	1	-0.05859	0.03531	-1.66	0.0972	-0.06144	8.38444
fundamental_r_ALPHABETICITY		1	0.00073269	0.00059797	1.23	0.2206	0.05206	11.04148

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-1.06936	6.66942
fundamental		1	0.89962	1.03079
r_ALPHABETICITY	Rank for Variable ALPHABETICITY	1	-0.12782	0.01065
fundamental_r_ALPHABETICITY		1	-0.00043979	0.00191

The interaction term for the Flesch Kincaid Reading Ease is not statistically significant.

Table 12 - Interaction effects model – Flesch Kincaid Reading Ease (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	1186417	395472	1078.70	<.0001
Error	2866	1050729	366.61879		
Corrected Total	2869	2237146			

Root MSE	19.14729	*R-Square*	0.5303
Dependent Mean	55.08118	*Adj R-Sq*	0.5298
Coeff Var	34.76195		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*
Intercept	Intercept	1	1.89306	1.90005	1.00	0.3192	0
fundamental		1	0.97278	0.03329	29.22	<.0001	0.70859
r_Flesch_K_R_E	Rank for Variable Flesch_Kincaid_Reading_Ease	1	-0.04173	0.03554	-1.17	0.2405	-0.04355
fundamental_r_Flesch_K_R_E		1	0.00059948	0.00060343	0.99	0.3206	0.04383

Parameter Estimates

Variable	*Label*	*DF*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	0	-1.83254	5.61865
fundamental		1	3.58797	0.90751	1.03805
r_Flesch_K_R_E	Rank for Variable Flesch_Kincaid_Reading_Ease	1	8.39973	-0.11142	0.02797
fundamental_r_Flesch_K_R_E		1	11.87814	-0.00058372	0.00178

The interaction term for the Flesch Kincaid Grade Level is also not statistically significant.

Table 13 - Interaction effects model – Flesch Kincaid Grade Level (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	1186603	395534	1079.06	<.0001
Error	2866	1050543	366.55384		
Corrected Total	2869	2237146			

Root MSE	19.14560	*R-Square*	0.5304
Dependent Mean	55.08118	*Adj R-Sq*	0.5299
Coeff Var	34.75887		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*
Intercept	Intercept	1	-2.32243	1.99619	-1.16	0.2448	0
fundamental		1	1.03381	0.03301	31.32	<.0001	0.75305
r_Flesch_K_G_L	Rank for Variable Flesch_Kincaid_Grade_Level	1	0.04817	0.03509	1.37	0.1700	0.05079
fundamental_r_Flesch_K_G_L		1	-0.00070391	0.00059752	-1.18	0.2389	-0.04756

Parameter Estimates

Variable	*Label*	*DF*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	0	-6.23655	1.59169
fundamental		1	3.52922	0.96908	1.09854
r_Flesch_K_G_L	Rank for Variable Flesch_Kincaid_Grade_Level	1	8.35497	-0.02064	0.11698
fundamental_r_Flesch_K_G_L		1	9.94946	-0.00188	0.00046771

Similarly, the interaction term for the Flesch Kincaid Grade Level is not statistically significant.

Table 14 - Interaction effects model – Gunning Fog Score (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	1186307	395436	1078.49	<.0001
Error	2866	1050839	366.65704		
Corrected Total	2869	2237146			

Root MSE	19.14829	*R-Square*	0.5303
Dependent Mean	55.08118	*Adj R-Sq*	0.5298
Coeff Var	34.76376		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	-1.20547	1.54954	-0.78	0.4367	0	0
fundamental		1	1.01950	0.02608	39.09	<.0001	0.74263	2.20233
r_Gunning_F_S	Rank for Variable Gunning_Fog_Score	1	0.03755	0.03465	1.08	0.2787	0.04006	8.34040
fundamental_r_Gunning_F_S		1	-0.00060656	0.00059045	-1.03	0.3044	-0.04022	9.35087

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-4.24381	1.83286
fundamental		1	0.96836	1.07064
r_Gunning_F_S	Rank for Variable Gunning_Fog_Score	1	-0.03040	0.10550
fundamental_r_Gunning_F_S		1	-0.00176	0.00055120

The interaction term for the Coleman Liau Index is not statistically significant.

Table 15 - Interaction effects model – Coleman Liau Index (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	1186775	395592	1079.40	<.0001
Error	2866	1050371	366.49371		
Corrected Total	2869	2237146			

Root MSE	19.14403	*R-Square*	0.5305
Dependent Mean	55.08118	*Adj R-Sq*	0.5300
Coeff Var	34.75602		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	-0.87823	2.01251	-0.44	0.6626	0	0
fundamental		1	0.99944	0.03349	29.85	<.0001	0.72801	3.63208
R_Coleman_L_I	Rank for Variable Coleman_Liau_Index	1	0.01626	0.03500	0.46	0.6423	0.01692	8.09485
fundamental_r_Coleman_L_I		1	0.00005600	0.00060003	0.09	0.9257	0.00376	9.90284

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-4.82434	3.06789
fundamental		1	0.93378	1.06510
R_Coleman_L_I	Rank for Variable Coleman_Liau_Index	1	-0.05237	0.08490
fundamental_r_Coleman_L_I		1	-0.00112	0.00123

The interaction term for the SMOG index is not statistically significant.

Table 16 - Interaction effects model – SMOG Index (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	1187046	395682	1079.92	<.0001
Error	2866	1050100	366.39917		
Corrected Total	2869	2237146			

Root MSE	19.14156	*R-Square*	0.5306
Dependent Mean	55.08118	*Adj R-Sq*	0.5301
Coeff Var	34.75153		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	-1.39854	1.54447	-0.91	0.3653	0	0
fundamental		1	1.02950	0.02567	40.10	<.0001	0.74991	2.13553
r_SMOG_Index	Rank for Variable SMOG_Index	1	0.04648	0.03530	1.32	0.1881	0.04841	8.25561
fundamental_r_SMOG_Index		1	-0.00099376	0.00060030	-1.66	0.0979	-0.06332	8.93337

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-4.42692	1.62983
fundamental		1	0.97916	1.07985
r_SMOG_Index	Rank for Variable SMOG_Index	1	-0.02275	0.11570
fundamental_r_SMOG_Index		1	-0.00217	0.00018330

The interaction term for the automated readability index is not statistically significant.

Table 17 - Interaction effects model – Automated Readability Index (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION
Dependent Variable: r_NO_OF_RECMNDTIONS Rank for Variable NO_OF_RECMNDTIONS

Number of Observations Read	2874
Number of Observations Used	2870
Number of Observations with Missing Values	4

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	1186768	395589	1079.38	<.0001
Error	2866	1050378	366.49632		
Corrected Total	2869	2237146			

Root MSE	19.14409	*R-Square*	0.5305
Dependent Mean	55.08118	*Adj R-Sq*	0.5300
Coeff Var	34.75614		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	-1.01371	2.01493	-0.50	0.6149	0	0
fundamental		1	1.00190	0.03348	29.92	<.0001	0.72981	3.63127
r_Auto_R_I	Rank for Variable Automated_Readability_Index	1	0.01891	0.03491	0.54	0.5881	0.01974	8.10660
fundamental_r_Auto_R_I		1	0.00000554	0.00059800	0.01	0.9926	0.00037285	9.88922

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-4.96456	2.93715
fundamental		1	0.93625	1.06756
r_Auto_R_I	Rank for Variable Automated_Readability_Index	1	-0.04954	0.08736
fundamental_r_Auto_R_I		1	-0.00117	0.00118

Of all the fluency measures tested on investment advisor data, a statistically significant interaction was only identified in the case of one – namely the corpus percentile. This is confirmed the F-ratio tests below which are only statistically significant in the case of the corpus percentile (see Appendix below).

5.6.3 Effect plots (Investment Data)

The effect plot below shows the change in the slope of the relationship between number of recommendations and percentile of financial fundamentals and visually demonstrates the presence of an interaction effect.

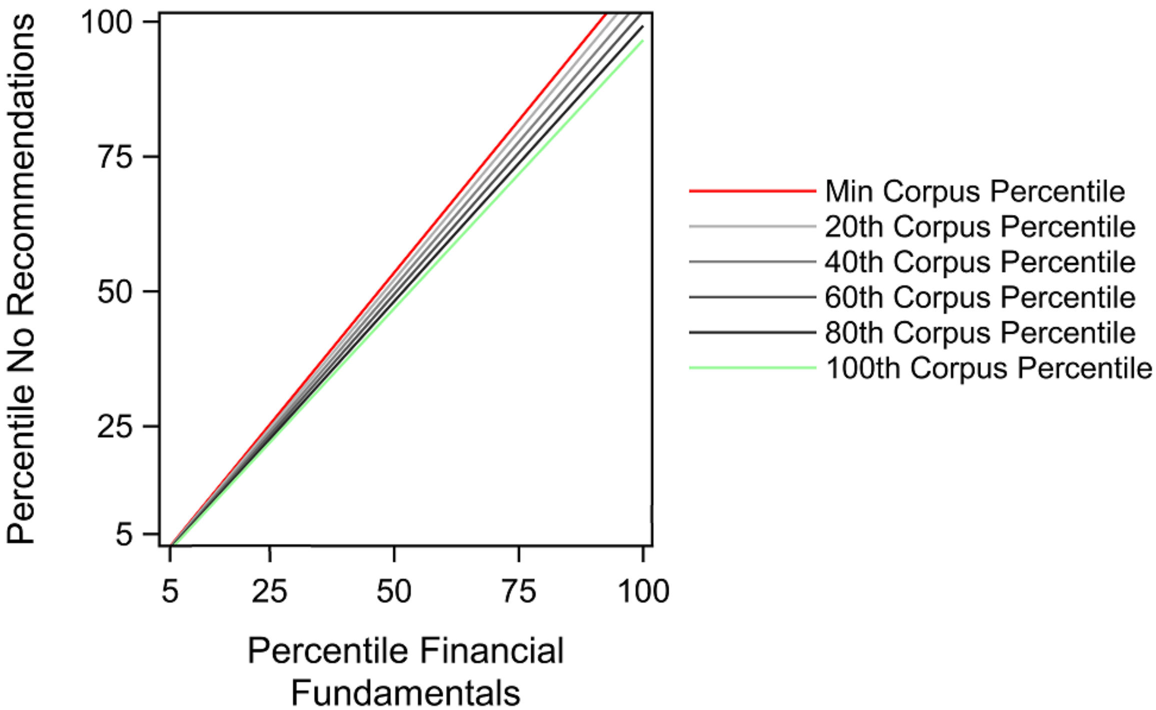


Figure 15 - Effect plot – Corpus Percentile (IAD)

It is of interest to note that whilst a statistically significant interaction was not found for the other fluency variables, the interaction is visible (admittedly) to a lesser degree in those variables too (see Appendix 4 below).

5.6.4 Coefficient plots - (Investment Data)

The coefficient plots bellows show the mean estimated regression coefficient for financial fundamentals together with the range around the mean at different levels of the interaction term. The exhibition of a clear trend or slope provides evidendence of interaction whilst a flat profile argues against the existence of an interaction.

The coefficient plot and spread of values (the horizontal lines at the top and bottom are minimum and maximum values) for corpus percentile exhibit the interaction effect between this measure of fluency and the number of recommendations provided by investment advisors.

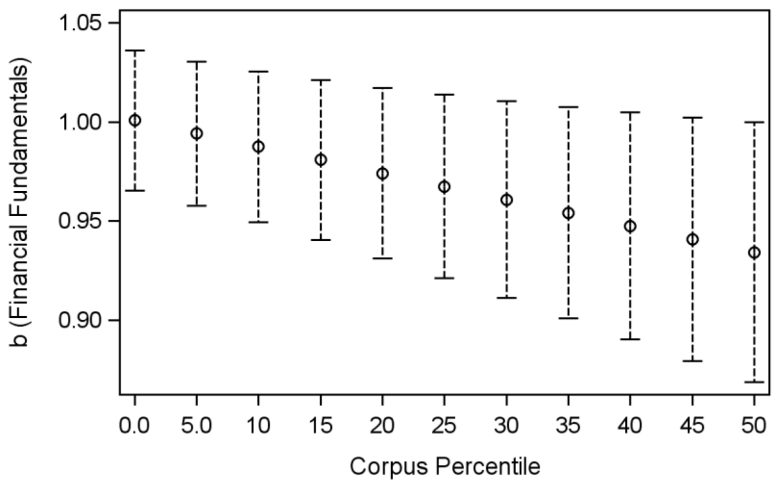


Figure 16 - Coefficient plot – Corpus Percentile (IAD)

It is also of interest to examine the coefficient plots for the variables that were not statistically significant moderators.

With the exception of the Gunning Fog Score and Automated Readability Index that show clearly flat lines on this plot, all the other do exhibit a slope however the scale of changed is very minor and therefore probably explains why the relationship was not statistically significant.

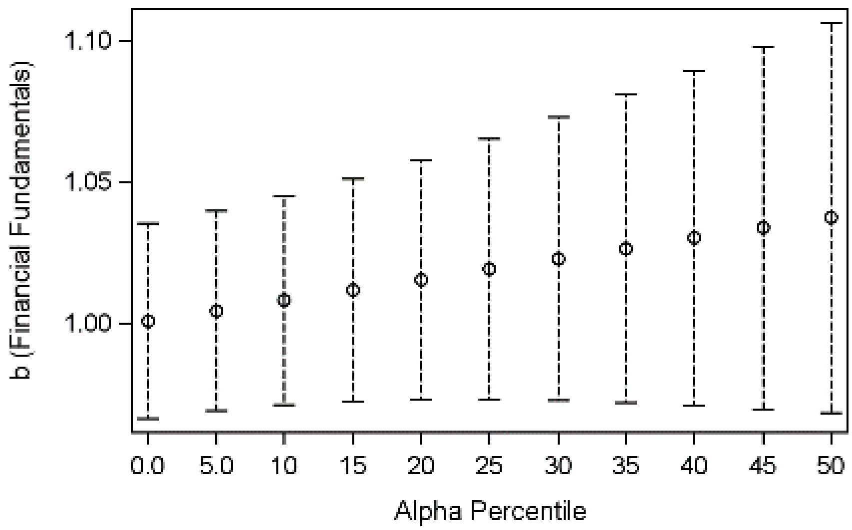


Figure 17 - Coefficient plot – Alphabeticity (IAD)

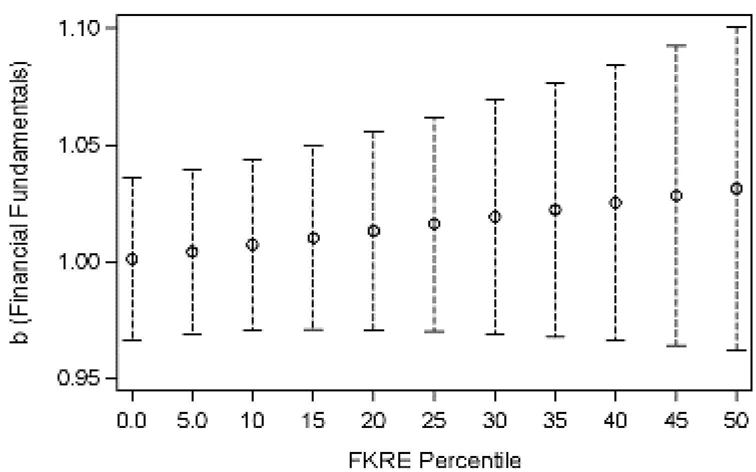


Figure 18 - Coefficient plot – Flesch Kincaid Reading Ease (IAD)

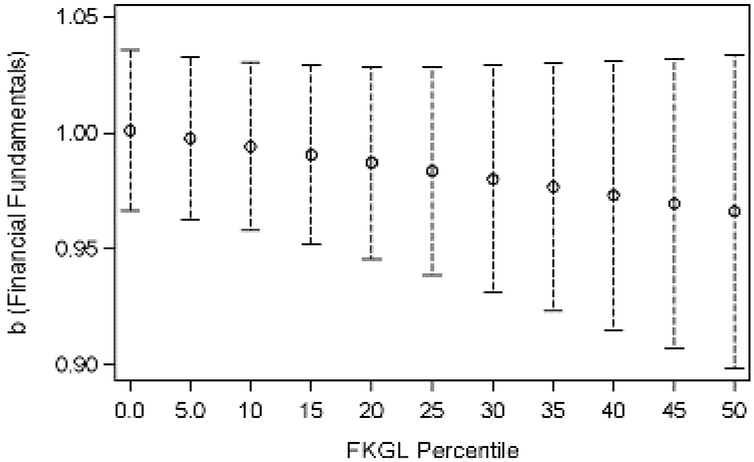


Figure 19 - Coefficient plot – Flesch Kincaid Grade Level (IAD)

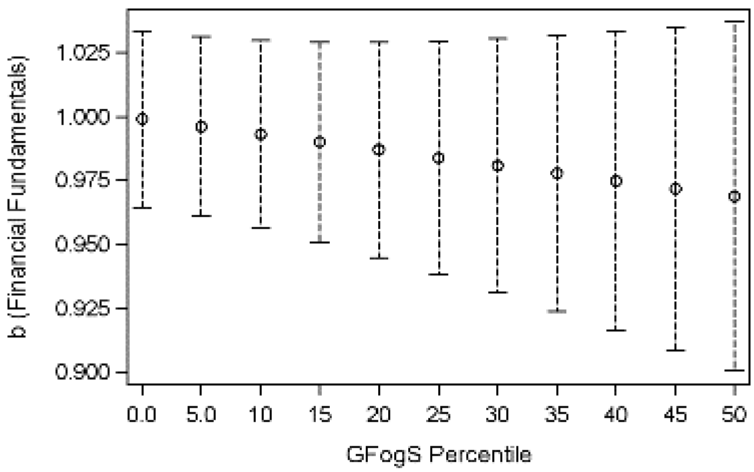


Figure 20 - Coefficient plot – Gunning Fog Score (IAD)

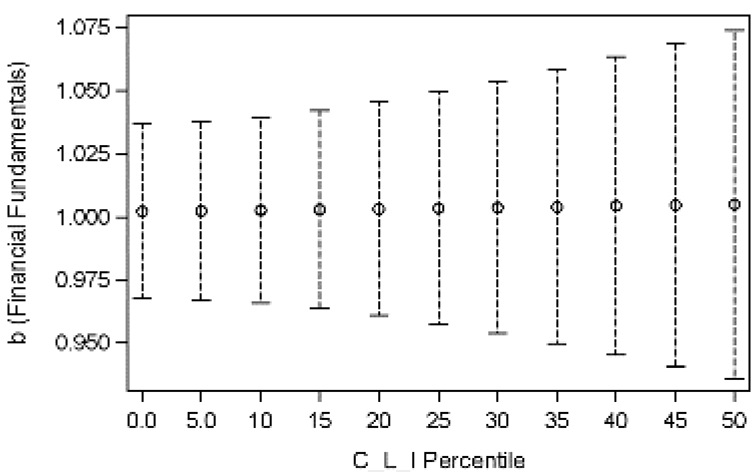


Figure 21 - Coefficient plot – Coleman Liau Index (IAD)

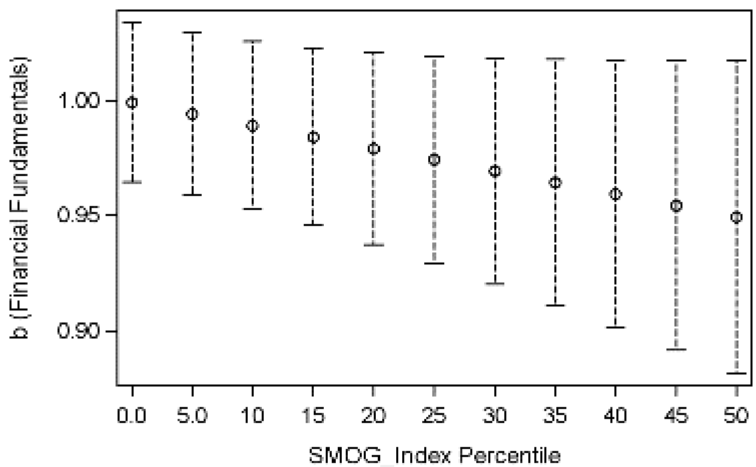


Figure 22 - Coefficient plot – SMOG Index (IAD)

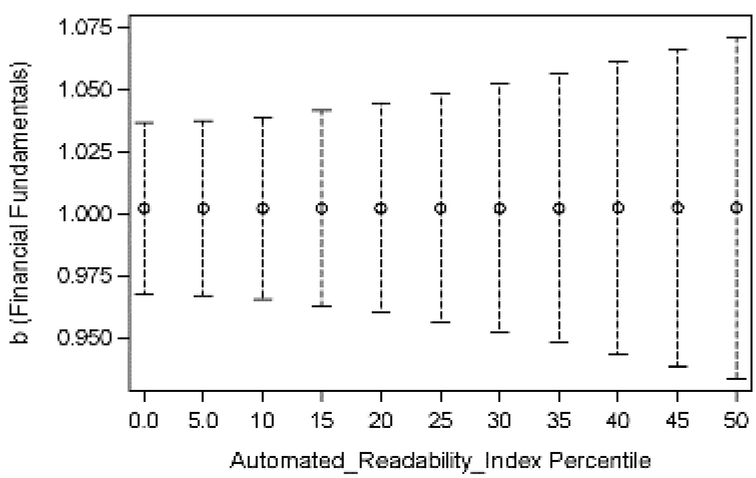
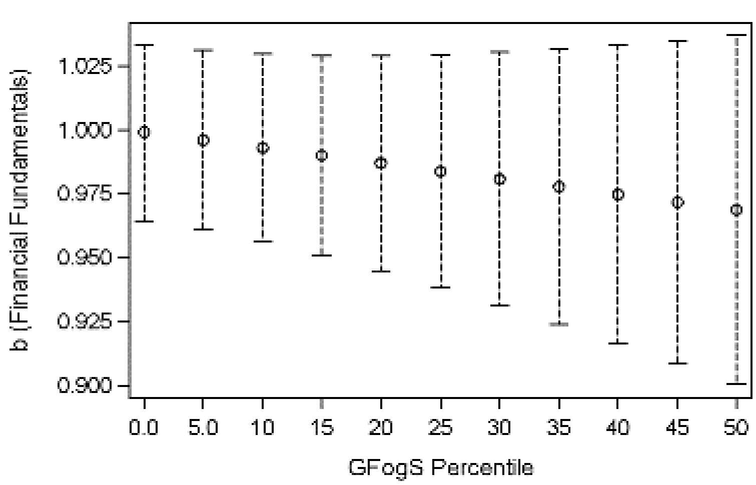


Figure 23 - Coefficient plot – Automated Readability Index (IAD)



5.7 Moderated Regression Analysis (Social Media Data)

The moderated regression models for the social media data are considered in this section.

5.7.1 Main effects model - (Social Media Data)

A remarkable and highly significant reversal occurs when we look at the main effect models for social media sentiment data. 7 out of the 8 measures of fluency calculated are statistically significant main effects in the model of fundamentals and fluency for social media data.

The corpus percentile is a statistically significant main effect.

Table 18 - Main effects model– Corpus Percentile (SMD)

INTERACTION AND MAIN EFFECT MODELS
CORPUS PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20044

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	7139366	3569683	7393.39	<.0001
Error	20041	9676216	482.82102		
Corrected Total	20043	16815582			

Root MSE	21.97319	*R-Square*	0.4246
Dependent Mean	49.35342	*Adj R-Sq*	0.4245
Coeff Var	44.52212		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	0.25738	0.46131	0.56	0.5769	0	0	-0.64683	1.16159
fundamental		1	0.99924	0.00824	121.31	<.0001	0.65103	1.00313	0.98309	1.01538
corpus_pctl		1	-0.00679	0.00409	-1.66	0.0964	-0.00892	1.00313	-0.01480	0.00122

Alphabeticity is a statistically significant main effect.

Table 19 - Main effects model – Alphabeticity (SMD)

INTERACTION AND MAIN EFFECT MODELS
Alphabeticity PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20044

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	7141171	3570586	7396.64	<.0001
Error	20041	9674411	482.73096		
Corrected Total	20043	16815582			

Root MSE	21.97114	*R-Square*	0.4247
Dependent Mean	49.35342	*Adj R-Sq*	0.4246
Coeff Var	44.51796		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	0.67286	0.50835	1.32	0.1856	0	0	-0.32355	1.66928
fundamental		1	0.99976	0.00822	121.56	<.0001	0.65137	1.00013	0.98364	1.01588
r_ALPHABETICITY	Rank for Variable ALPHABETICITY	1	-0.01359	0.00533	-2.55	0.0108	-0.01366	1.00013	-0.02404	-0.00315

The Flesch Kincaid Reading Ease is a statistically significant main effect.

Table 20 - Main effects model – Flesch Kincaid Reading Ease (SMD)

INTERACTION AND MAIN EFFECT MODELS
FKRE PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	7125953	3562977	7386.04	<.0001
Error	20001	9648346	482.39316		
Corrected Total	20003	16774299			

Root MSE	21.96345	*R-Square*	0.4248
Dependent Mean	49.39562	*Adj R-Sq*	0.4248
Coeff Var	44.46437		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	1.51732	0.49996	3.03	0.0024	0	0
fundamental		1	1.00030	0.00823	121.48	<.0001	0.65155	1.00031
r_Flesch_K_R_E	Rank for Variable Flesch_Kincaid_Reading_Ease	1	-0.03259	0.00540	-6.03	<.0001	-0.03235	1.00031

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	0.53735	2.49728
fundamental		1	0.98416	1.01644
r_Flesch_K_R_E	Rank for Variable Flesch_Kincaid_Reading_Ease	1	-0.04318	-0.02200

The Flesch Kincaid Grade Level is a statistically significant main effect

Table 21 - Main effects model – Flesch Kincaid Grade Level (SMD

INTERACTION AND MAIN EFFECT MODELS
FKGL PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	7122409	3561205	7379.66	<.0001
Error	20001	9651890	482.57037		
Corrected Total	20003	16774299			

Root MSE	21.96748	*R-Square*	0.4246
Dependent Mean	49.39562	*Adj R-Sq*	0.4245
Coeff Var	44.47253		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	-1.35838	0.50626	-2.68	0.0073	0	0
fundamental		1	1.00031	0.00824	121.45	<.0001	0.65155	1.00040
r_Flesch_K_G_L	Rank for Variable Flesch_Kincaid_Grade_Level	1	0.02883	0.00535	5.39	<.0001	0.02890	1.00040

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-2.35069	-0.36606
fundamental		1	0.98416	1.01645
r_Flesch_K_G_L	Rank for Variable Flesch_Kincaid_Grade_Level	1	0.01834	0.03932

The Gunning Fog Score is a statistically significant main effect.

Table 22 - Main effects model – Gunning Fog Score (SMD)

INTERACTION AND MAIN EFFECT MODELS
GFS PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	7190372	3595186	7502.91	<.0001
Error	20001	9583927	479.17237		
Corrected Total	20003	16774299			

Root MSE	21.89001	*R-Square*	0.4287
Dependent Mean	49.39562	*Adj R-Sq*	0.4286
Coeff Var	44.31568		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	2.99954	0.48951	6.13	<.0001	0	0	2.04005	3.95902
fundamental		1	0.98708	0.00826	119.51	<.0001	0.64294	1.01322	0.97089	1.00327
r_Gunning_F_S	Rank for Variable Gunning_Fog_Score	1	-0.06661	0.00509	-13.08	<.0001	-0.07037	1.01322	-0.07660	-0.05663

The Coleman Liau Index is a statistically significant main effect

Table 23 Main effects model – Coleman Liau Index (SMD)

INTERACTION AND MAIN EFFECT MODELS
CLI PERCENTILE- StockTwits and Twitter

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	7111038	3555519	7359.21	<.0001
Error	20001	9663261	483.13889		
Corrected Total	20003	16774299			

Root MSE	21.98042	*R-Square*	0.4239
Dependent Mean	49.39562	*Adj R-Sq*	0.4239
Coeff Var	44.49872		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	-0.60407	0.51411	-1.17	0.2400	0	0	-1.61176	0.40362
fundamental		1	0.99998	0.00824	121.31	<.0001	0.65134	1.00084	0.98382	1.01614
R_Coleman_L_I	Rank for Variable Coleman_Liau_Index	1	0.01265	0.00542	2.34	0.0195	0.01254	1.00084	0.00204	0.02327

Table 24 - Main effects model – SMOG Index (SMD)

INTERACTION AND MAIN EFFECT MODELS
SMI PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	7111330	3555665	7359.73	<.0001
Error	20001	9662969	483.12430		
Corrected Total	20003	16774299			

Root MSE	21.98009	*R-Square*	0.4239
Dependent Mean	49.39562	*Adj R-Sq*	0.4239
Coeff Var	44.49805		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	0.51193	0.47695	1.07	0.2831	0	0	-0.42293	1.44680
fundamental		1	0.99822	0.00825	120.94	<.0001	0.65019	1.00352	0.98204	1.01440
r_SMOG_Index	Rank for Variable SMOG_Index	1	-0.01324	0.00538	-2.46	0.0138	-0.01324	1.00352	-0.02377	-0.00270

Table 25 - Main effects model – Automated Readability Index (SMD)

INTERACTION AND MAIN EFFECT MODELS
ARI PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	7109450	3554725	7356.35	<.0001
Error	20001	9664849	483.21830		
Corrected Total	20003	16774299			

Root MSE	21.98223	*R-Square*	0.4238
Dependent Mean	49.39562	*Adj R-Sq*	0.4238
Coeff Var	44.50238		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	-0.37382	0.51609	-0.72	0.4689	0	0
fundamental		1	0.99985	0.00825	121.26	<.0001	0.65125	1.00124
r_Auto_R_I	Rank for Variable Automated_Readability_Index	1	0.00796	0.00540	1.47	0.1408	0.00791	1.00124

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-1.38540	0.63777
fundamental		1	0.98369	1.01601
r_Auto_R_I	Rank for Variable Automated_Readability_Index	1	-0.00263	0.01855

5.7.2 Interaction Effects Model - (Social Media Data)

The interaction and moderation effects are reviewed in this section. Whilst 7 out of the 8 fluency measures were significant main effects - once the model is fully specified the following variables are found to have a statistically interaction (moderating effect) in a model of fundamental performance and social media sentiment:

- The corpus percentile
- The Gunning Fog Score
- The Coleman Liau Index
- The SMOG Index; and
- The Automated Readability index

The Corpus percentile exhibits statistically significant interaction.

Table 26 - Interaction Effects Model – Corpus Percentile (SMD)

INTERACTION AND MAIN EFFECT MODELS
CORPUS PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20044

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7141661	2380554	4931.43	<.0001
Error	20040	9673922	482.73062		
Corrected Total	20043	16815582			

Root MSE	21.97113	R-Square	0.4247
Dependent Mean	49.35342	Adj R-Sq	0.4246
Coeff Var	44.51795		

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	1	0.93967	0.55742	1.69	0.0919	0	0
fundamental		1	0.98533	0.01042	94.60	<.0001	0.64197	1.60429
corpus_pctl		1	-0.03162	0.01210	-2.61	0.0090	-0.04153	8.79527
fundamental_corpus_pctl		1	0.00051844	0.00023780	2.18	0.0293	0.03534	9.15161

Parameter Estimates				
Variable	Label	DF	95% Confidence Limits	
Intercept	Intercept	1	-0.15292	2.03225
fundamental		1	0.96492	1.00575
corpus_pctl		1	-0.05534	-0.00791
fundamental_corpus_pctl		1	0.00005232	0.00098455

Alphabeticity is not a statistically significant interaction variable.

Table 27 - Interaction effects model – Alphabeticity (SMD)

INTERACTION AND MAIN EFFECT MODELS
Alphabeticity PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20044

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	7142179	2380726	4932.06	<.0001
Error	20040	9673403	482.70475		
Corrected Total	20043	16815582			

Root MSE	21.97054	*R-Square*	0.4247
Dependent Mean	49.35342	*Adj R-Sq*	0.4246
Coeff Var	44.51676		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	1.66503	0.85430	1.95	0.0513	0	0
fundamental		1	0.97990	0.01602	61.17	<.0001	0.63843	3.79466
r_ALPHABETICITY	Rank for Variable ALPHABETICITY	1	-0.03423	0.01525	-2.25	0.0248	-0.03442	8.18630
fundamental_r_ALPHABETICITY		1	0.00041442	0.00028679	1.45	0.1485	0.02554	10.87889

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-0.00947	3.33952
fundamental		1	0.94850	1.01130
r_ALPHABETICITY	Rank for Variable ALPHABETICITY	1	-0.06412	-0.00435
fundamental_r_ALPHABETICITY		1	-0.00014771	0.00097656

The Flesch Kincaid Reading Ease is not a statistically significant interaction variable.

Table 28 - Interaction effects model – Flesch Kincaid Reading Ease (SMD)

INTERACTION AND MAIN EFFECT MODELS
FKRE PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	7125958	2375319	4923.79	<.0001
Error	20000	9648341	482.41704		
Corrected Total	20003	16774299			

Root MSE	21.96399	*R-Square*	0.4248
Dependent Mean	49.39562	*Adj R-Sq*	0.4247
Coeff Var	44.46547		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*
Intercept	Intercept	1	1.58806	0.86362	1.84	0.0660	0
fundamental		1	0.99890	0.01619	61.68	<.0001	0.65063
r_Flesch_K_R_E	Rank for Variable Flesch_Kincaid_Reading_Ease	1	-0.03409	0.01588	-2.15	0.0318	-0.03384
fundamental_r_Flesch_K_R_E		1	0.00002957	0.00029432	0.10	0.9200	0.00184

Parameter Estimates

Variable	*Label*	*DF*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	0	-0.10471	3.28084
fundamental		1	3.86848	0.96716	1.03064
r_Flesch_K_R_E	Rank for Variable Flesch_Kincaid_Reading_Ease	1	8.63542	-0.06521	-0.00297
fundamental_r_Flesch_K_R_E		1	11.66857	-0.00054732	0.00060646

The Flesch Kincaid Grade Level is not a statistically significant interaction variable.

Table 29 - Interaction effects model – Flesch Kincaid Grade Level (SMD)

INTERACTION AND MAIN EFFECT MODELS
FKGL PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	7122449	2374150	4919.57	<.0001
Error	20000	9651850	482.59251		
Corrected Total	20003	16774299			

Root MSE	21.96799	*R-Square*	0.4246
Dependent Mean	49.39562	*Adj R-Sq*	0.4245
Coeff Var	44.47355		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*
Intercept	Intercept	1	-1.55516	0.85263	-1.82	0.0682	0
fundamental		1	1.00416	0.01576	63.70	<.0001	0.65406
r_Flesch_K_G_L	Rank for Variable Flesch_Kincaid_Grade_Level	1	0.03308	0.01575	2.10	0.0357	0.03316
fundamental_r_Flesch_K_G_L		1	-0.00008364	0.00029161	-0.29	0.7742	-0.00514

Parameter Estimates

Variable	*Label*	*DF*	*Variance Inflation*	*95% Confidence Limits*	
Intercept	Intercept	1	0	-3.22639	0.11606
fundamental		1	3.66479	0.97326	1.03506
r_Flesch_K_G_L	Rank for Variable Flesch_Kincaid_Grade_Level	1	8.65893	0.00222	0.06395
fundamental_r_Flesch_K_G_L		1	11.14263	-0.00065522	0.00048793

The Gunning Fog Score is a statistically significant interaction variable.

Table 30 - Interaction effects model – Gunning Fog Score (SMD)

INTERACTION AND MAIN EFFECT MODELS
GFS PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	7202891	2400964	5016.95	<.0001
Error	20000	9571408	478.57040		
Corrected Total	20003	16774299			

Root MSE	21.87625	*R-Square*	0.4294
Dependent Mean	49.39562	*Adj R-Sq*	0.4293
Coeff Var	44.28784		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	0.49264	0.69251	0.71	0.4769	0	0
fundamental		1	1.03652	0.01271	81.55	<.0001	0.67514	2.40234
r_Gunning_F_S	Rank for Variable Gunning_Fog_Score	1	0.00149	0.01426	0.10	0.9167	0.00158	7.94819
fundamental_r_Gunning_F_S		1	-0.00139	0.00027257	-5.11	<.0001	-0.08018	8.61513

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-0.86474	1.85002
fundamental		1	1.01160	1.06143
r_Gunning_F_S	Rank for Variable Gunning_Fog_Score	1	-0.02645	0.02943
fundamental_r_Gunning_F_S		1	-0.00193	-0.00085980

The Coleman Liau Index is a statistically significant interaction variable.

Table 31 - Interaction effects model – Coleman Liau Index (SMD)

INTERACTION AND MAIN EFFECT MODELS
CLI PERCENTILE- StockTwits and Twitter

The REG Procedure
Model: INTERACTION
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	7113965	2371322	4909.40	<.0001
Error	20000	9660334	483.01668		
Corrected Total	20003	16774299			

Root MSE	21.97764	*R-Square*	0.4241
Dependent Mean	49.39562	*Adj R-Sq*	0.4240
Coeff Var	44.49309		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	1.14587	0.87722	1.31	0.1915	0	0
fundamental		1	0.96562	0.01621	59.57	<.0001	0.62896	3.87128
R_Coleman_L_I	Rank for Variable Coleman_Liau_Index	1	-0.02390	0.01580	-1.51	0.1305	-0.02369	8.52368
fundamental_r_Coleman_L_I		1	0.00072257	0.00029351	2.46	0.0138	0.04406	11.12374

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-0.57355	2.86529
fundamental		1	0.93385	0.99739
R_Coleman_L_I	Rank for Variable Coleman_Liau_Index	1	-0.05487	0.00708
fundamental_r_Coleman_L_I		1	0.00014727	0.00130

The SMOG index is a statistically significant interaction variable.

Table 32 - Interaction effects model – SMOG Index (SMD)

INTERACTION AND MAIN EFFECT MODELS
SMI PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	7117546	2372515	4913.69	<.0001
Error	20000	9656753	482.83766		
Corrected Total	20003	16774299			

Root MSE	21.97357	*R-Square*	0.4243
Dependent Mean	49.39562	*Adj R-Sq*	0.4242
Coeff Var	44.48485		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	-1.12510	0.65993	-1.70	0.0882	0	0
fundamental		1	1.03058	0.01222	84.31	<.0001	0.67127	2.20257
r_SMOG_Index	Rank for Variable SMOG_Index	1	0.03718	0.01504	2.47	0.0135	0.03718	7.86245
fundamental_r_SMOG_Index		1	-0.00102	0.00028346	-3.59	0.0003	-0.05684	8.71837

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-2.41863	0.16842
fundamental		1	1.00662	1.05454
r_SMOG_Index	Rank for Variable SMOG_Index	1	0.00769	0.06667
fundamental_r_SMOG_Index		1	-0.00157	-0.00046146

The SMOG index is a statistically significant interaction variable.

Table 33 - Interaction effects model – Automated Readability Index (SMD)

INTERACTION AND MAIN EFFECT MODELS
ARI PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION
Dependent Variable: r_TOTAL_BULLISH Rank for Variable TOTAL_BULLISH

Number of Observations Read	20044
Number of Observations Used	20004
Number of Observations with Missing Values	40

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	7112448	2370816	4907.58	<.0001
Error	20000	9661851	483.09256		
Corrected Total	20003	16774299			

Root MSE	21.97937	*R-Square*	0.4240
Dependent Mean	49.39562	*Adj R-Sq*	0.4239
Coeff Var	44.49659		

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Standardized Estimate*	*Variance Inflation*
Intercept	Intercept	1	1.40962	0.88250	1.60	0.1102	0	0
fundamental		1	0.96492	0.01626	59.33	<.0001	0.62850	3.89674
r_Auto_R_I	Rank for Variable Automated_Readability_Index	1	-0.02900	0.01579	-1.84	0.0663	-0.02882	8.54800
fundamental_r_Auto_R_I		1	0.00072975	0.00029294	2.49	0.0127	0.04457	11.11365

Parameter Estimates

Variable	*Label*	*DF*	*95% Confidence Limits*	
Intercept	Intercept	1	-0.32015	3.13940
fundamental		1	0.93304	0.99680
r_Auto_R_I	Rank for Variable Automated_Readability_Index	1	-0.05995	0.00195
fundamental_r_Auto_R_I		1	0.00015557	0.00130

5.7.3 F-Ratio test - (Social Media Data)

The F-Ratio test is a further marker of statistically significant interactions and confirms below the findings of the main and interaction effects regression models.

A statistically significant interaction will have a P-Value less than 0.05 as displayed below for all statistically significant interaction terms which are highlighted.

Table 34 - F-Ratio test – Corpus Percentile (SMD)

INTERACTION AND MAIN EFFECT MODELS
CORPUS PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_TOTAL_BULLISH

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	2294.32108	4.75	0.0293
Denominator	20040	482.73062		

Table 35 - F-Ratio test – Alphabeticity (SMD)

INTERACTION AND MAIN EFFECT MODELS
Alphabeticity PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_TOTAL_BULLISH

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	1007.96149	2.09	0.1485
Denominator	20040	482.70475		

Table 36 - F-Ratio test – Flesch Kincaid Reading Ease (SMD)

INTERACTION AND MAIN EFFECT MODELS
FKRE PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_TOTAL_BULLISH

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	4.86911	0.01	0.9200
Denominator	20000	482.41704		

Table 37- F-Ratio test – Flesch Kincaid Grade Level (SMD)

INTERACTION AND MAIN EFFECT MODELS
FKGL PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_TOTAL_BULLISH				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	39.70614	0.08	0.7742
Denominator	20000	482.59251		

Table 38 - F-Ratio test – Gunning Fog Score (SMD)

INTERACTION AND MAIN EFFECT MODELS
GFS PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_TOTAL_BULLISH				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	12519	26.16	<.0001
Denominator	20000	478.57040		

Table 39- F-Ratio test – Coleman Liau Index (SMD)

INTERACTION AND MAIN EFFECT MODELS
CLI PERCENTILE- StockTwits and Twitter

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_TOTAL_BULLISH				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	2927.38036	6.06	0.0138
Denominator	20000	483.01668		

Table 40 - F-Ratio test – SMOG Index (SMD)

INTERACTION AND MAIN EFFECT MODELS
SMI PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_TOTAL_BULLISH

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	6216.05939	12.87	0.0003
Denominator	20000	482.83766		

Table 41 - F-Ratio test – Automated Readability Index (SMD)

INTERACTION AND MAIN EFFECT MODELS
ARI PERCENTILE - StockTwits and Twitter

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_TOTAL_BULLISH

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	2997.98987	6.21	0.0127
Denominator	20000	483.09256		

5.7.4 Coefficient plot - (Social Media Data)

The extent of interaction is clearly visible in the coefficient plots displayed below, especially in the case of the Gunning Fog Score.

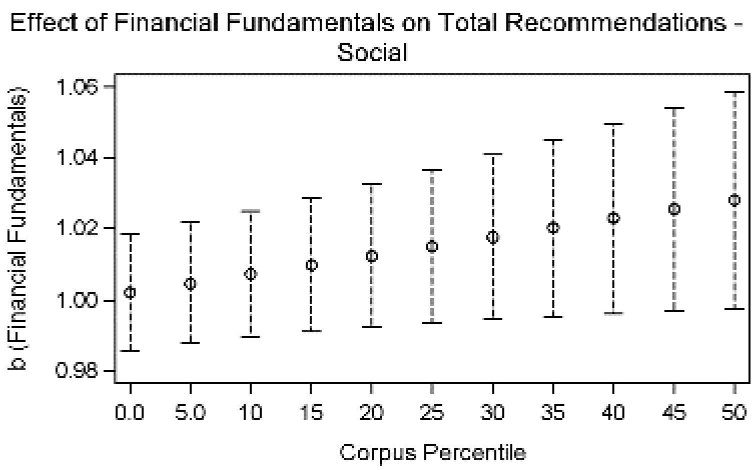


Figure 24 -- Coefficient plot – Corpus Percentile (SMD)

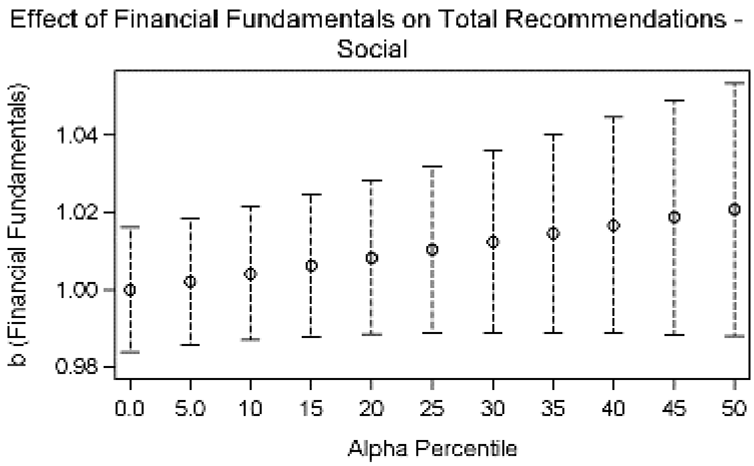


Figure 25 - Coefficient plot – Alphabeticity (SMD)

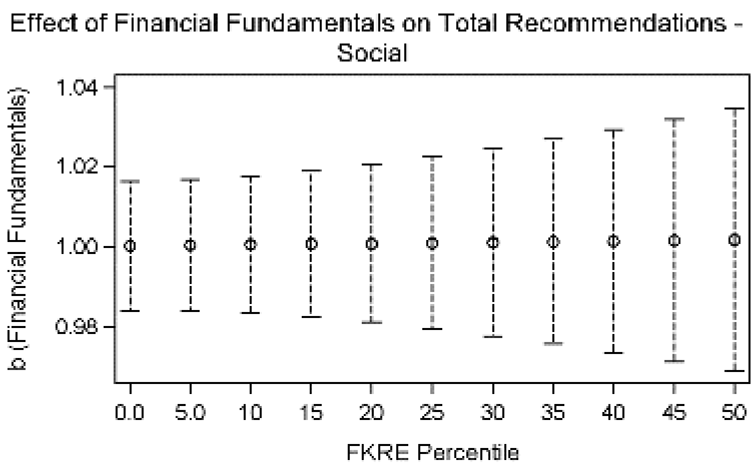


Figure 26 - Coefficient plot – Flesch Kincaid Reading Ease (SMD)

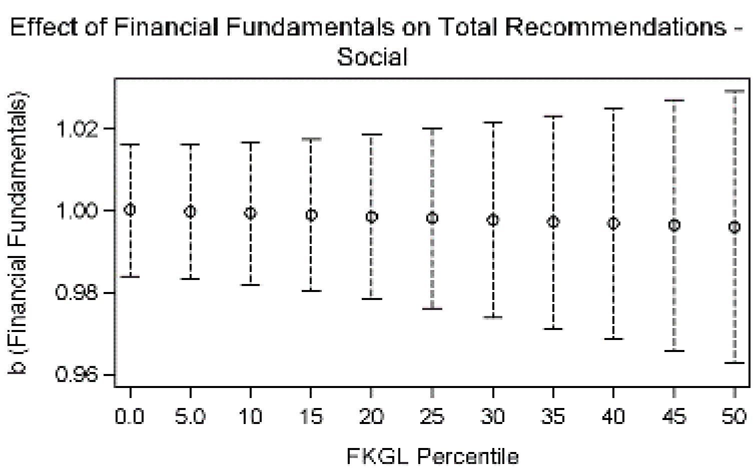


Figure 27 - Coefficient plot – Flesch Kincaid Grade Level (SMD)

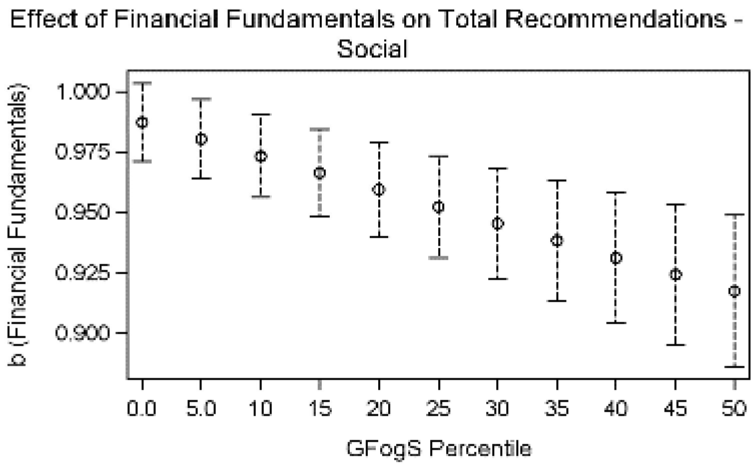


Figure 28 - Coefficient plot – Gunning Fog Score (SMD)

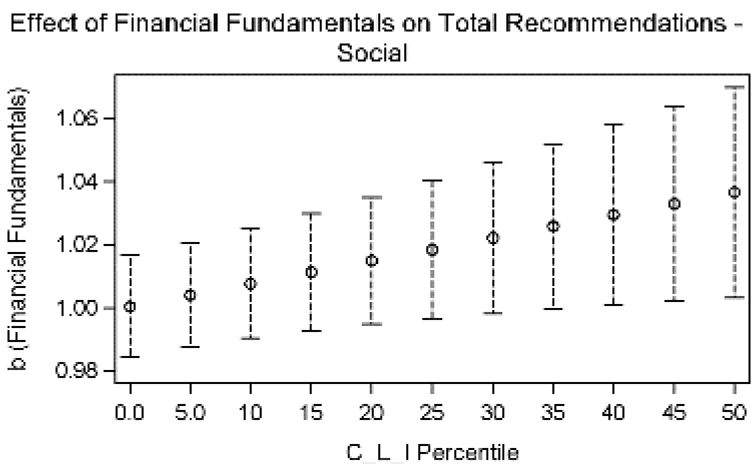


Figure 29 - Coefficient plot – Coleman Liau Index (SMD)

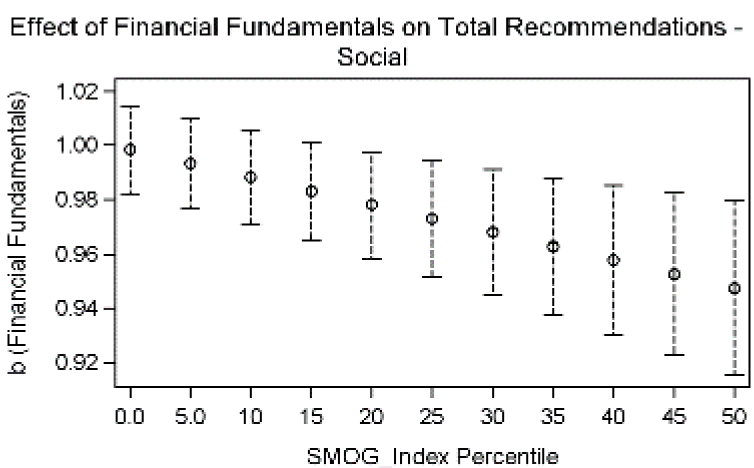


Figure 30 - Coefficient plot – SMOG Index (SMD)

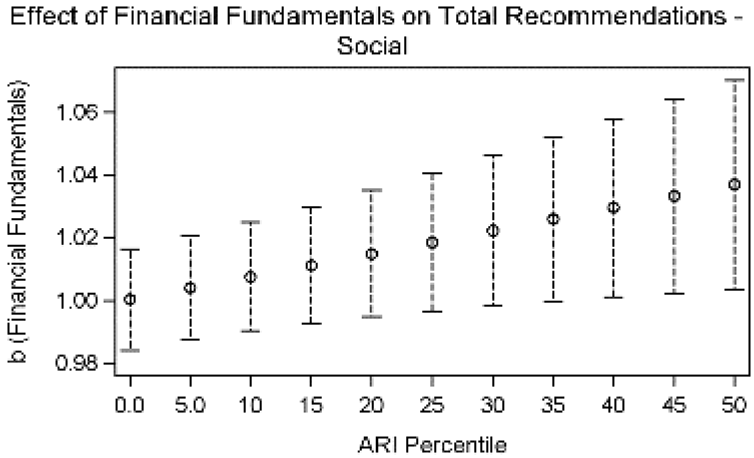


Figure 31 - Coefficient plot – Automated Readability Index (SMD)

6 Discussion of Research Results

6.1 Introduction and summary of results

The following findings are made following the data analysis

Research Question	Tests Performed	Section Reference	Finding
1. The diversity of opinion obtained from social media differs from that obtained from investment experts 2. The concentration of opinion and investment advice differs between social media sources	Distribution analysis	• 5.4	Investment advice is the most centralised and non-diverse, followed by StockTwits which are generated by a semi-closed community, with the most diversity of opinion found in the unfiltered Twitter feed
3. The fluency of corporate name is a mediator of the effect of fundamental performance on investment analyst sentiment, and the effect differs for the different measures of fluency	Distribution analysis Moderated regression analysis	• 5.4 • 5.6	Some evidence found for one out of 8 measures of fluency
4. The fluency of corporate name is a mediator of the effect of fundamental performance on social media sentiment, and the effect differs for different sources of social media sentiment data	Distribution analysis	• 5.4 • 5.7	Some channels (e.g. StockTwits) were much less easily predicted by the model and specific measures i.e. The Gunning Fog Score showed markedly superior power in distribution analysis. The moderated regression analysis indicated that Social Media data are markedly more influenced by fluency than investment advisor sentiments.

6.2 Discussion of results

The following key points are noted following the analysis that was performed in this research:

- The diagnostic analysis for the main effects models for fluency and fundamental performance (for both investment advisors and in social media sentiment) showed that the model was most predictive at identifying the higher percentile values of sentiment (higher rankings in number of recommendations) than it was at determining lower values. This was a key finding, which was not anticipated in the literature reviewed.

The main research questions, 3.1 and 4.1 were stated as follows:
Research question 3.1

- The fluency of corporate name is a mediator of the effect of fundamental performance on investment analyst sentiment, and the effect differs for the different measures of fluency

Research question 4.1

- The fluency of corporate name is a mediator of the effect of fundamental performance on social media sentiment, and the effect differs for different sources of social media sentiment data

The justification for testing these questions was found in the increasing application of crowd sourced intelligence in serious applications which has been cited in several sources that refer to the wisdom of the crowds ((Chen, De, Hu, & Hwang, 2014; Franch, 2013; Sarasohn-Kahn, 2008; Surowiecki, 2005).

Partial support for this view was found in the distribution analysis of sentiment, which showed a marked rank ordering of sentiment in terms diversity from crowd to expert – with the expert sources showing the least heterogeneity of opinion, and the open or crowd sources (the unfiltered twitter stock database) providing the best variety.

However when the actual moderated regression model for the effect of sentiment was tested on the investment advisor data, the analysis showed that out of all 8 fluency measures tested as possible main effects in a joint model to predict investor advisor sentiment with financial fundamentals, only the corpus percentile was a statistically significant main effect. The corpus percentile was also the only statistically significant moderator for investment advisor data.

This finding is reversed in a significant and remarkable manner when the social media data is considered. In this case we find that 7 out of the 8 measures of fluency that are tested are statistically significant.

This research therefore makes a significant and remarkable finding that directly contradicts the theory of the wisdom of the crowds, namely that social media stock sentiment data are significantly more moderated by the spurious effect of fluency of corporate name than are data generated by investment advisors and closed communities of experts.

7 Conclusion

7.1 Key Research Findings

The research conducted makes a key finding that shows **empirically** that social media stock sentiment data are more influenced by the spurious effect of fluency of corporate name than are stock sentiments generated by traditional expert communities. This finding flies in the face of an increasingly adopted thesis that postulates superior intelligence can be obtained from an aggregation of unstructured crowd-sourced data.

The research conducted also makes a key finding that the model of sentiment does not predict equally well at all levels of sentiment – models are better at identifying highly recommended equities than they are at identifying those that are not

The research also makes a further finding that the different measures of fluency have different levels of effectiveness at predicting sentiment levels.

7.2 Recommendations for further research

During the course of this research a painstaking amount of effort was undertaken to make a close statistical examination of an open question and ensure that valid quantitative contributions (however narrow in focus) could be made to the understanding of an increasingly important subject area.

To this end data were collated from a variety of sources and a lot of effort expended to generate multiple measures of fluency, to transform the data and to undertake rigorous quantitative analysis in the pursuit of relevant insights.

The diagnostic results provide rich suggestions of additional investigation that is required to improve understanding of the topic. It is of interest to gain a deeper understanding of the reasons why the models specified work better at higher levels of sentiment and to understand the underlying drivers of the influential observations and outliers observed in the model diagnostics.

It is also of interest to perform additional analysis to understand why the corpus percentile out of all the fluency measures turns out to be a statistically significant main effect. It was the only fluency measure, which did show that it had an interaction effect on the opinion of experts which in itself is a significant finding. These two areas provide rich ground for meaningful further research.

8 References

References

Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proceedings of the National Academy of Sciences of the United States of America, 103*(24), 9369-9372. doi:0601071103 [pii]

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research [review article]. *Computational Intelligence Magazine, IEEE, 9*(2), 48-57.

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

Corredor, P., Ferrer, E., & Santamaria, R. (2014). Is cognitive bias really present in analyst forecasts? the role of investor sentiment. *International Business Review, 23*(4), 824-837.

Fiske, S. T., & Taylor, S. E. (1991). Social cognition, 2nd. *NY: McGraw-Hill,* , 16-15.

Foroudi, P., Melewar, T. C., & Gupta, S. (2014). Linking corporate logo, corporate image, and reputation: An examination of consumer perceptions in the financial setting. *Journal of Business Research, 67*(11), 2269-2281. doi:http://dx.doi.org/10.1016/j.jbusres.2014.06.015

Franch, F. (2013). (Wisdom of the crowds) 2: 2010 UK election prediction with social media. *Journal of Information Technology & Politics, 10*(1), 57-71.

Graf, L. K., & Landwehr, J. R. (2015). A dual-process perspective on fluency-based aesthetics: The pleasure-interest model of aesthetic liking. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc,* doi:1088868315574978 [pii]

Green, T. C., & Jame, R. (2013). Company name fluency, investor recognition, and firm value. *Journal of Financial Economics, 109*(3), 813-834.

Hacıbedel, B. (2014). Does investor recognition matter for asset pricing? *Emerging Markets Review, 21*, 1-20.

Hern, A. (2015,). Why google is restructuring, why the name alphabet and how it affects you. *The Guardian* doi:http://www.theguardian.com/technology/2015/aug/11/google-alphabet-why-change-restructuring-what-it-means

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science (New York, N.Y.), 349*(6245), 261-266. doi:10.1126/science.aaa8685 [doi]

Hong, H., & Kacperczyk, M. (2009). The price of sin: The effects of social norms on markets. *Journal of Financial Economics, 93*(1), 15-36.

Huang, D., Jiang, F., Tu, J., & Zhou, G. (2014). Investor sentiment aligned: A powerful predictor of stock returns. *Review of Financial Studies,* , hhu080.

Itzkowitz, J., Itzkowitz, J., & Rothbort, S. (2015). ABCs of trading: Behavioral biases affect stock turnover and value. *Review of Finance,* , rfv012.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697.

Khataei, A., & Arya, A. (2015). Trust-based individualization for persuasive presentation builder. *Human-computer interaction: Interaction technologies* (pp. 418-428) Springer.

Kim, S., Koh, Y., Cha, J., & Lee, S. (2015). Effects of social media on firm value for US restaurant companies. *International Journal of Hospitality Management, 49*, 40-46.

Kleinnijenhuis, J., Schultz, F., Oegema, D., & van Atteveldt, W. (2013). Financial news and market panics in the age of high-frequency sentiment trading algorithms. *Journalism,* , 1464884912468375.

Korfiatis, N., García-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications, 11*(3), 205-217.

Lick, D. J., & Johnson, K. L. (2015). The interpersonal consequences of processing ease fluency as a metacognitive foundation for prejudice. *Current Directions in Psychological Science, 24*(2), 143-148.

Manke, S. N., & Shivale, N. (2015). A review on: Opinion mining and sentiment analysis based on natural language processing. *International Journal of Computer Applications, 109*(4)

Nardo, M., Petracco-Giudici, M., & Naltsidis, M. (2015). WALKING DOWN WALL STREET WITH A TABLET: A SURVEY OF STOCK MARKET PREDICTIONS USING THE WEB. *Journal of Economic Surveys,*

Oliveira, N., Cortez, P., & Areal, N. (2013). On the predictability of stock market behavior using stocktwits sentiment and posting volume. *Progress in artificial intelligence* (pp. 355-365) Springer.

Olson, E. L., & Thjømøe, H. M. (2009). Sponsorship effect metric: Assessing the financial value of sponsoring by comparisons to television advertising. *Journal of the Academy of Marketing Science, 37*(4), 504-515.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Rushe, D., & Sam, T. (2015,). **Google to restructure into new holding company called alphabet**
. *The Guardian* doi:http://www.theguardian.com/technology/2015/aug/10/google-alphabet-parent-company

Sarasohn-Kahn, J. (2008). *The Wisdom of Patients: Health Care Meets Online Social Media,*

Segaran, T., & Hammerbacher, J. (2009). *Beautiful data: The stories behind elegant data solutions* " O'Reilly Media, Inc.".

Srinivasan, R., & Umashankar, N. (2014). There's something in a name: Value relevance of congruent ticker symbols. *Customer Needs and Solutions, 1*(3), 241-252.

Statistical Analysis System Institute. (1999). *SAS/STAT user's guide, version 8* SAS Institute.

Stevenson, A. (2010). *Oxford dictionary of english* Oxford University Press.

Surowiecki, J. (2005). *The wisdom of crowds* Anchor.

Tan, H., Wang, E., & Zhou, B. (2013). When the use of positive language backfires: The joint effect of language sentiment, readability, and investor sophistication on earnings judgments. *Journal of Accounting Research, Forthcoming,*

Wei, Y., Song, P., & Rutherford, B. (2014). Corporate blogs, social media links and firm performance: A study of fortune 500 companies. *Journal of Digital & Social Media Marketing, 2*(2), 159-175.

9 Consistency Matrix

R1: The diversity of opinion obtained from social media differs from that obtained from investment experts	All references in Chapter 2. Purely quantitative analysis.
R2: The concentration of opinion and investment advice differs between social media sources	All references in Chapter 2. Purely quantitative analysis.
R3: The fluency of corporate name is a mediator of the effect of fundamental performance on investment analyst sentiment	All references in Chapter 2. Purely quantitative analysis.
R4: The fluency of corporate name is a mediator of the effect of fundamental performance on social media sentiment	All references in Chapter 2. Purely quantitative analysis.

10 Appendix 1 – Diagnostic Plots (Investment Advisor Data)

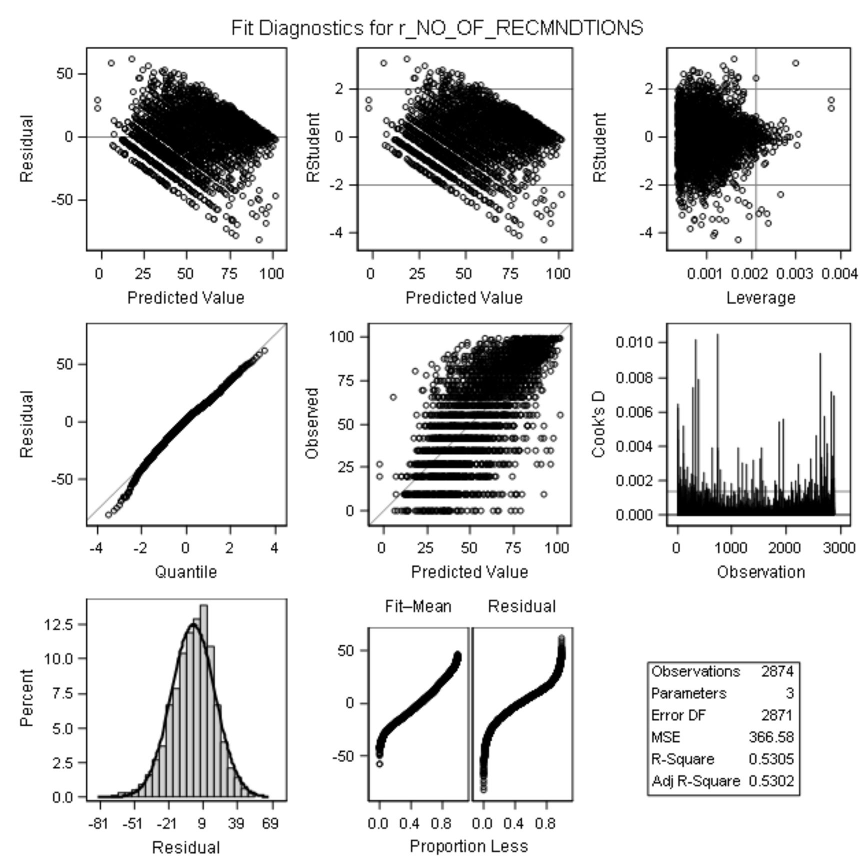


Figure 32 - Diagnostic Plots - Alphabeticity (IAD)

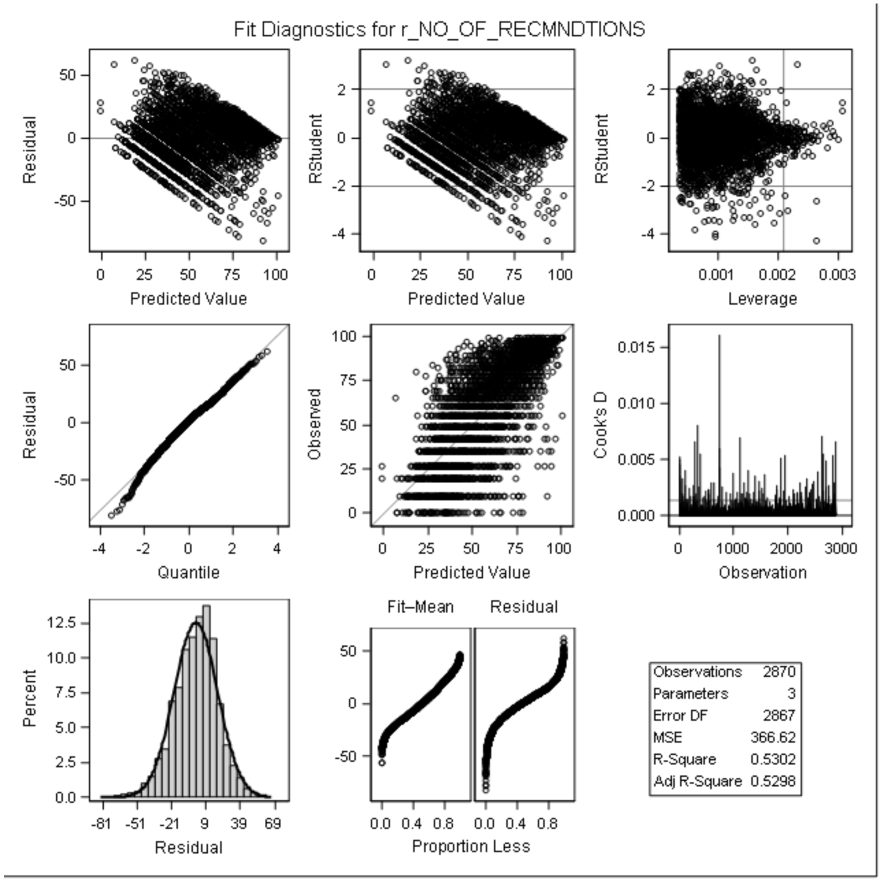


Figure 33 - Diagnostic Plots - Flesch Kincaid Reading Ease (IAD)

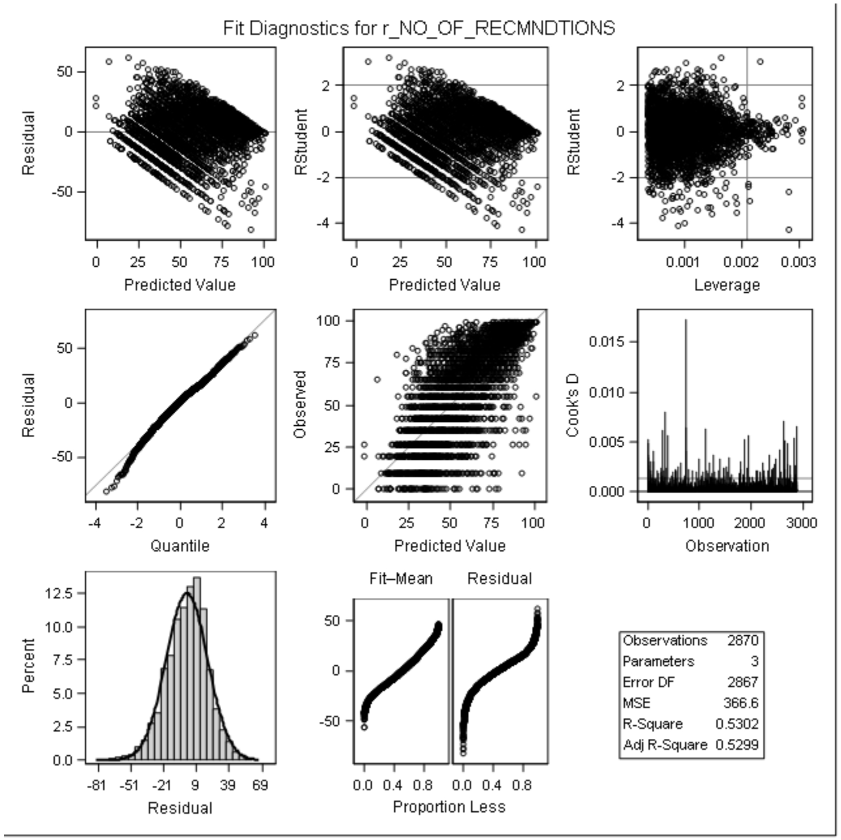


Figure 34 - Diagnostic Plots - Flesch Kincaid Grade Level (IAD)

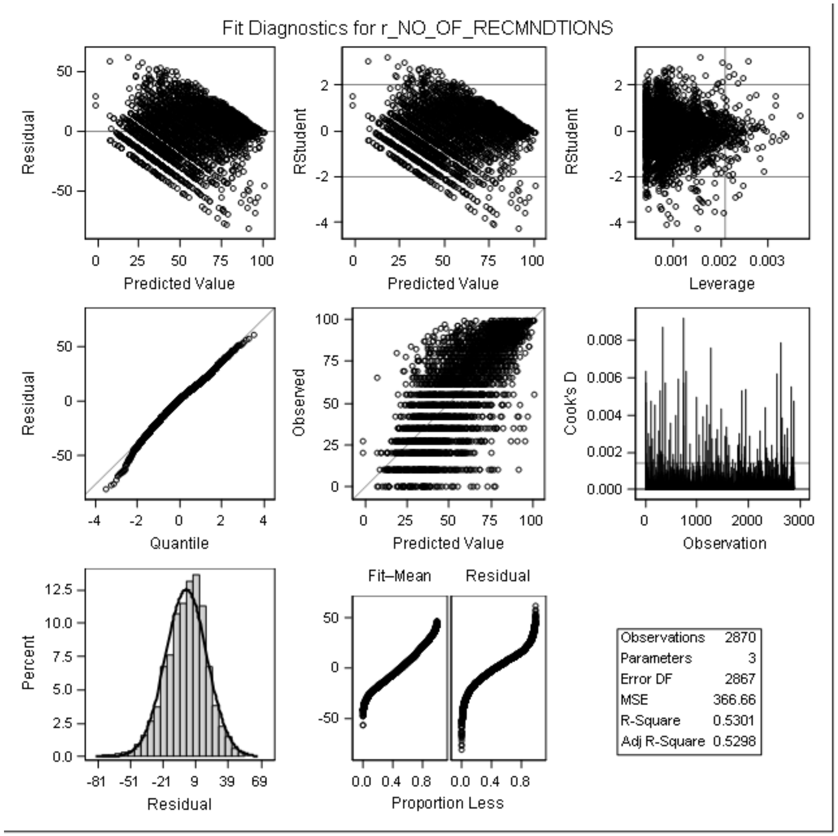


Figure 35 - Diagnostic Plots - Gunning Fog Score (IAD)

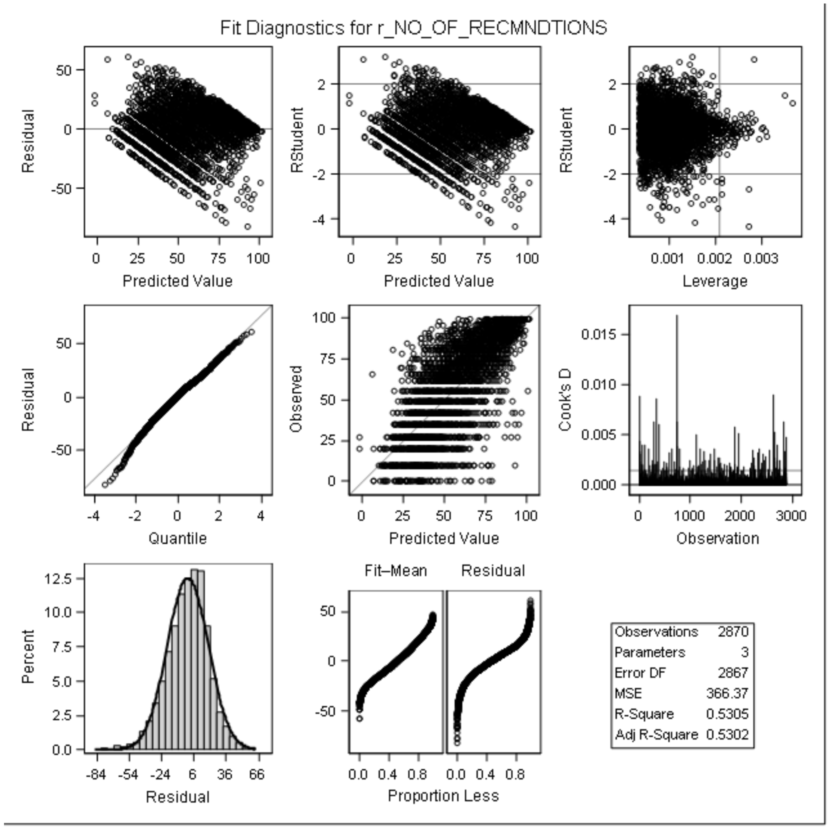


Figure 36 - Diagnostic Plots - Coleman Liau Index (IAD)

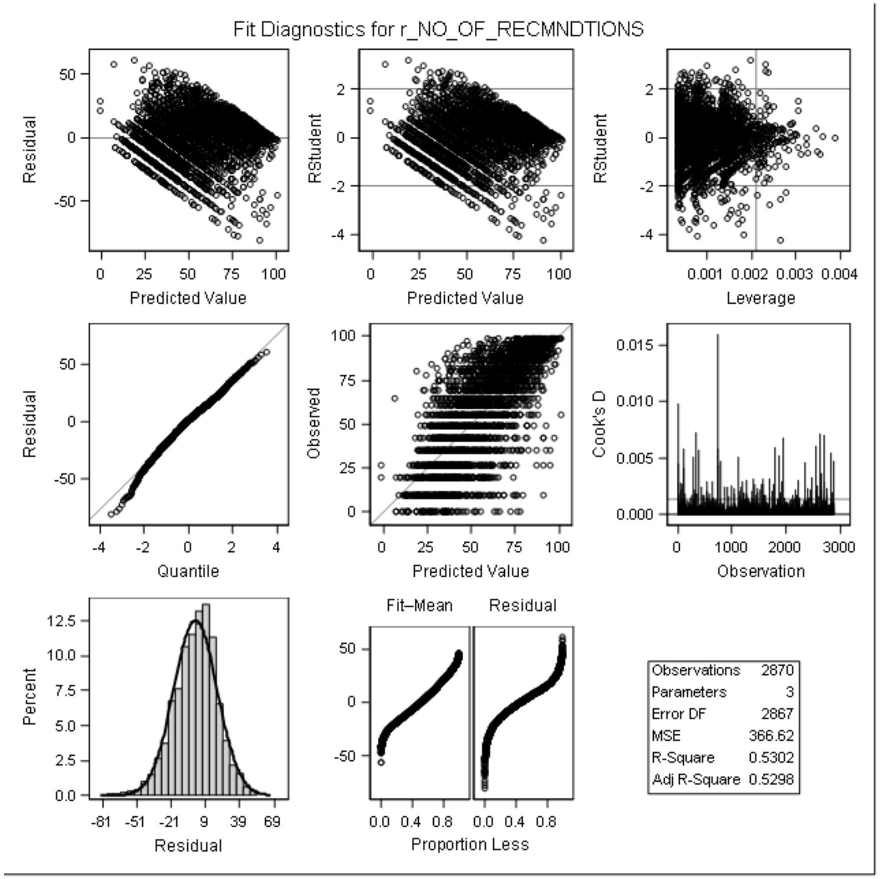


Figure 37 - Diagnostic Plots - SMOG Index (IAD)

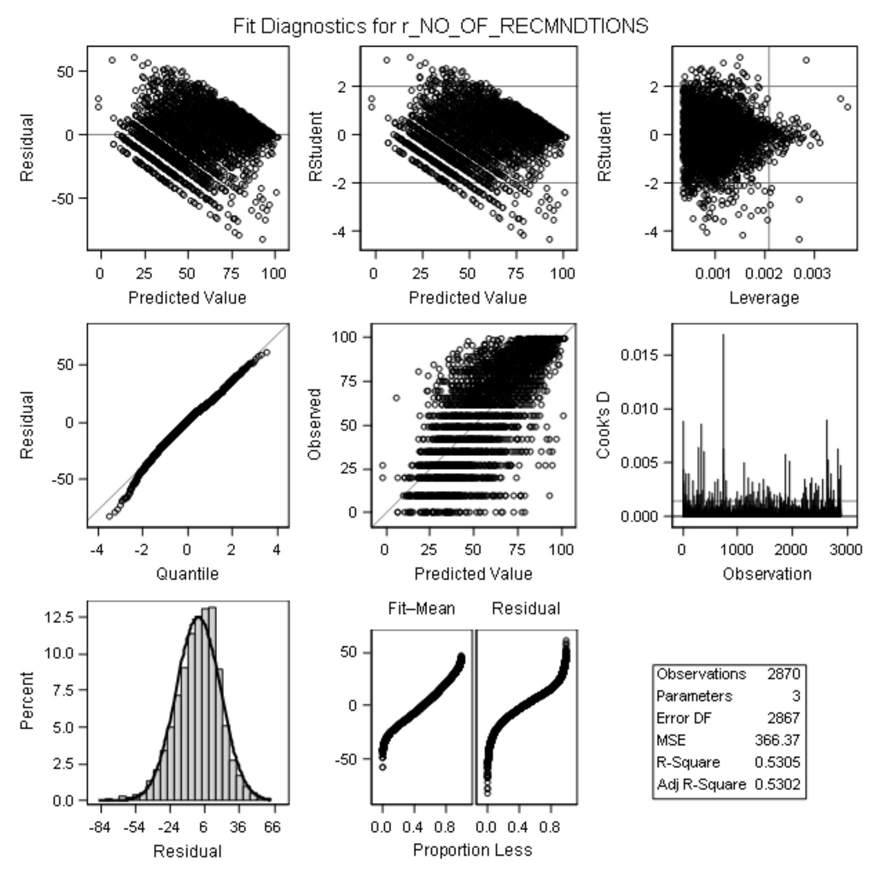


Figure 38 - Diagnostic Plots - Automated Readability Index (IAD)

11 Appendix 2 – Diagnostic plots social media data

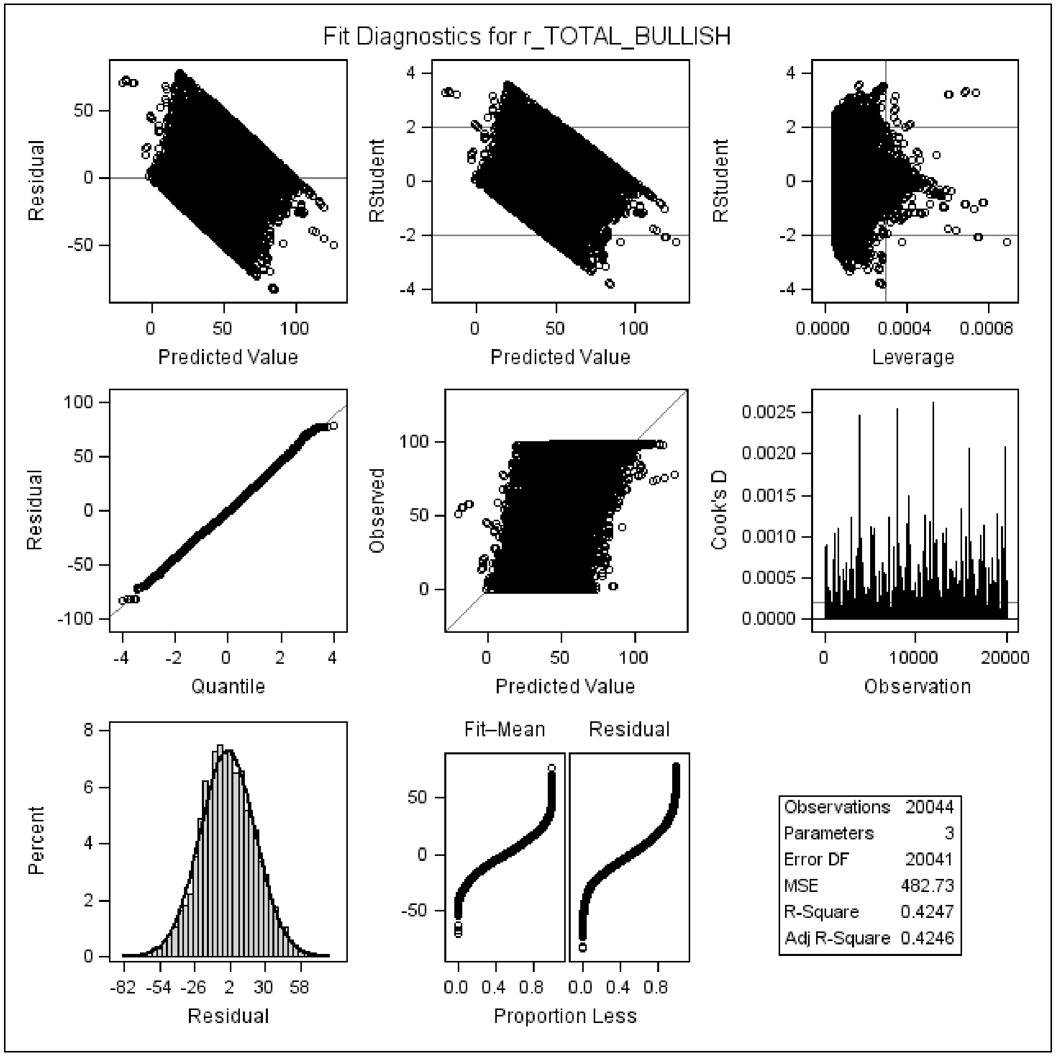


Figure 39 - Diagnostic Plots – Alphabeticity SMD

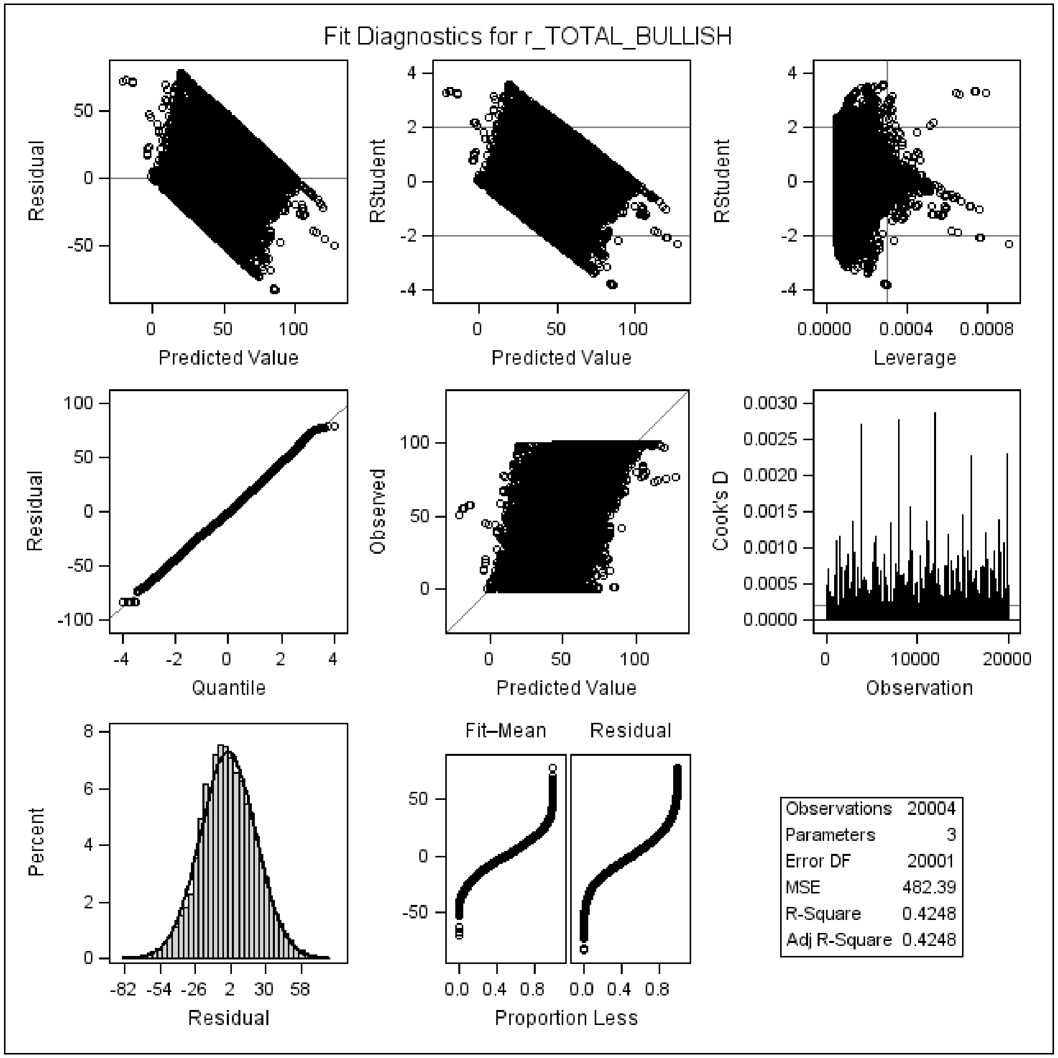


Figure 40 - Diagnostic Plots - Flesch Kincaid Reading Ease SMD

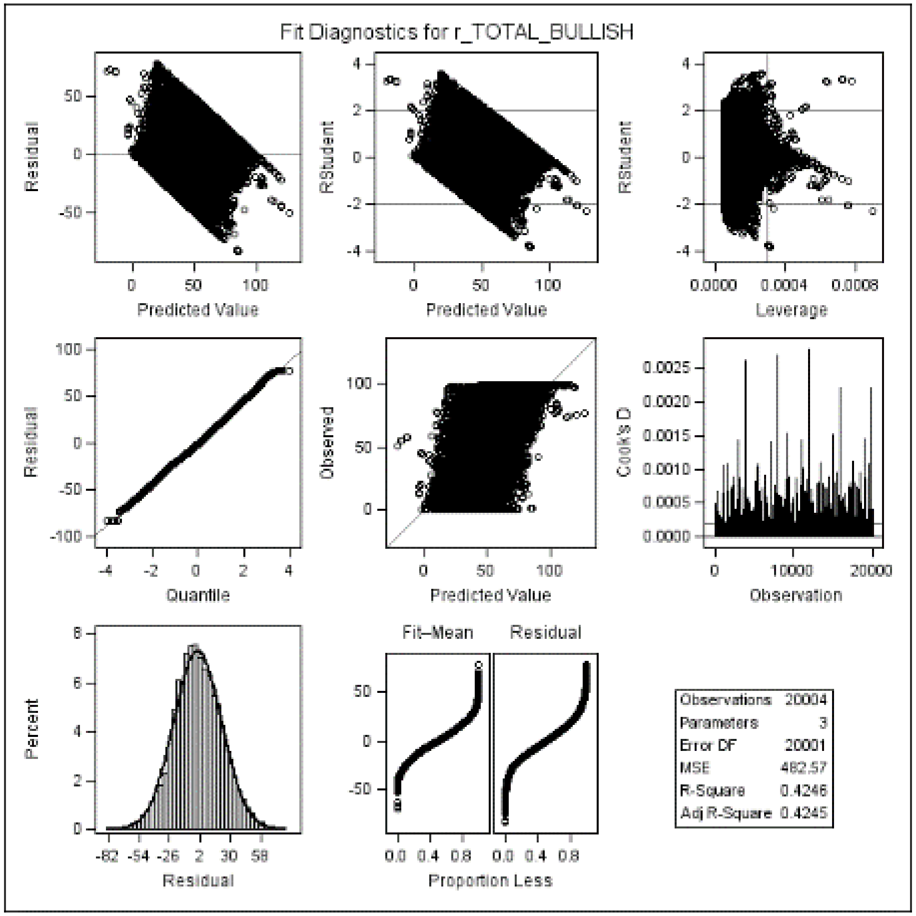


Figure 41 - Diagnostic Plots - Flesch Kincaid Grade Level SMD

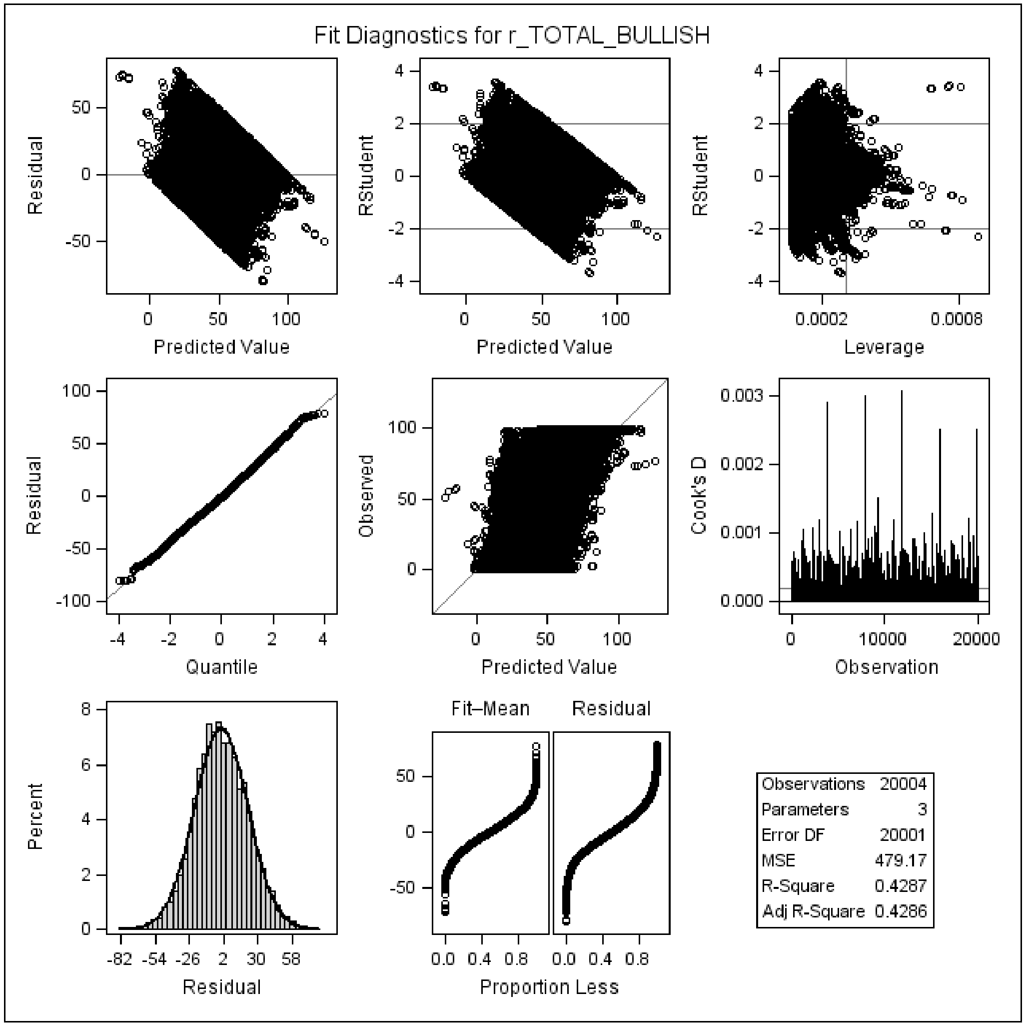


Figure 42 - Diagnostic Plots - Gunning Fog Score SMD

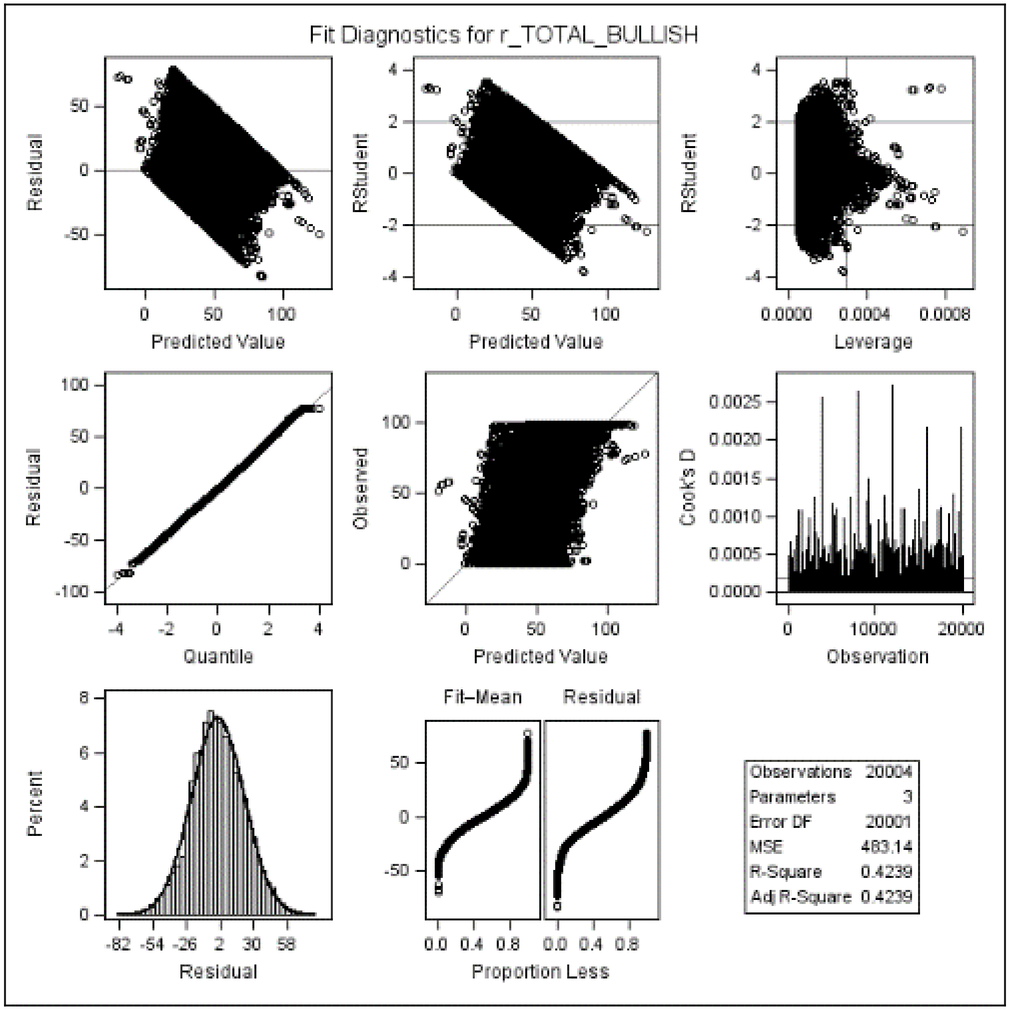


Figure 43 - Diagnostic Plots - Coleman Liau Index SMD

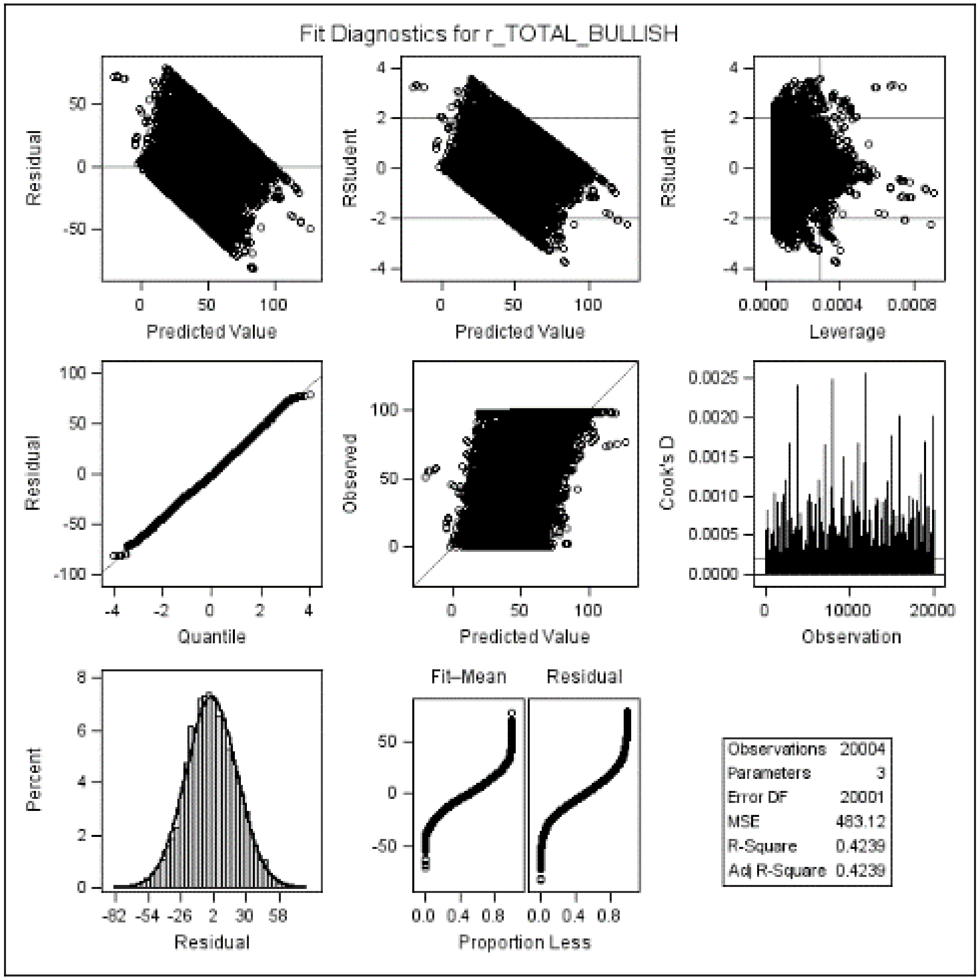


Figure 44 - Diagnostic Plots - SMOG Index SMD

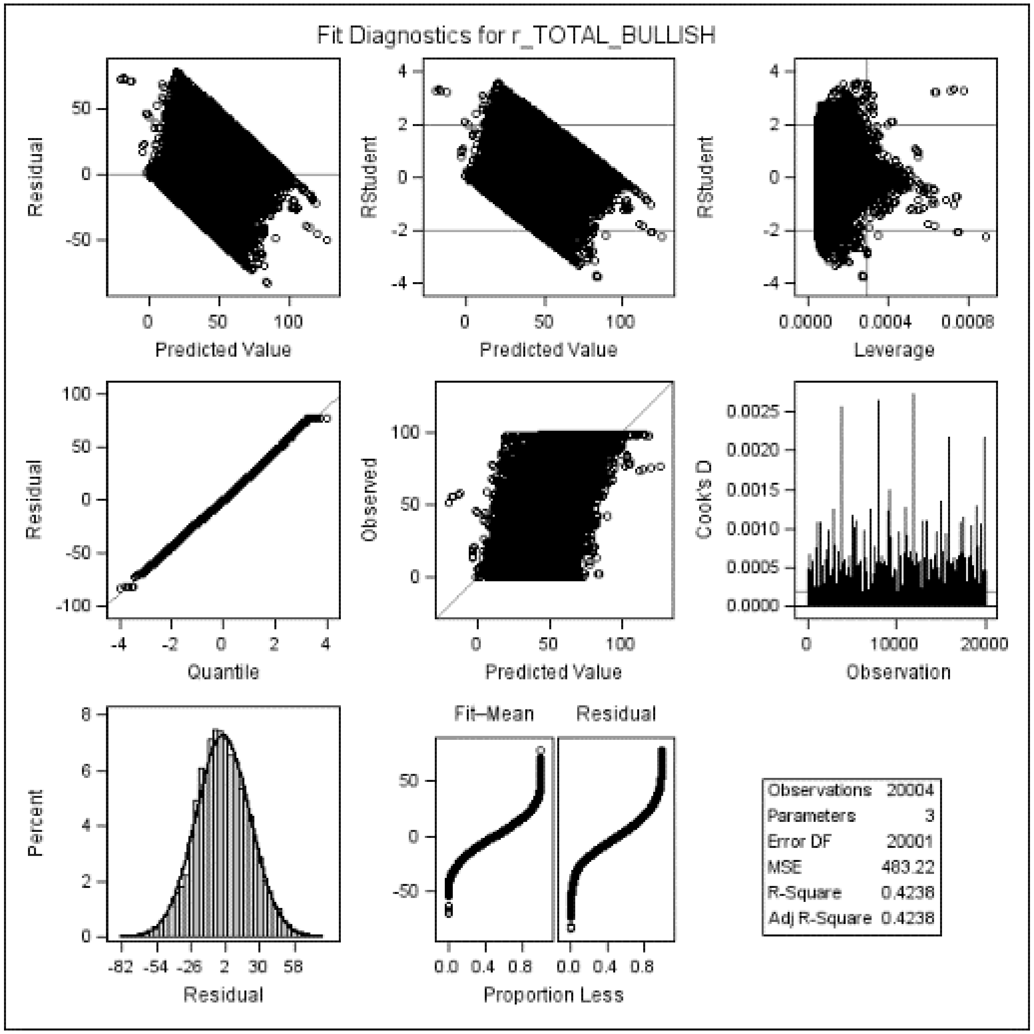


Figure 45 - Diagnostic Plots - Automated Readability Index SMD

Appendix 3 – F-Ratio Tests Investment Advisor Data

Table 42 - F-Ratio test – Corpus Percentile (IAD)

Test INT_EFFECT Results for Dependent Variable r_NO_OF_RECMNDTIONS

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	2425.90938	6.63	0.0101
Denominator	2870	365.64957		

Table 43 - F-Ratio test – Alphabeticity (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_NO_OF_RECMNDTIONS

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	550.27698	1.50	0.2206
Denominator	2870	366.51341		

Table 44 - F-Ratio test – Flesch Kincaid Reading Ease (IAD)

__Effect of Financial Fundamentals on Total Recommendations__

__The REG Procedure__
__Model: INTERACTION__

Test INT_EFFECT Results for Dependent Variable r_NO_OF_RECMNDTIONS				
Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	361.83823	0.99	0.3206
Denominator	2866	366.61879		

Table 45 - F-Ratio test – Flesch Kincaid Grade Level (IAD)

__Effect of Financial Fundamentals on Total Recommendations__

__The REG Procedure__
__Model: INTERACTION__

Test INT_EFFECT Results for Dependent Variable r_NO_OF_RECMNDTIONS				
Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	508.69824	1.39	0.2389
Denominator	2866	366.55384		

Table 46 - F-Ratio test – Gunning Fog Score (IAD)

__Effect of Financial Fundamentals on Total Recommendations__

__The REG Procedure__
__Model: INTERACTION__

Test INT_EFFECT Results for Dependent Variable r_NO_OF_RECMNDTIONS				
Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	386.93369	1.06	0.3044
Denominator	2866	366.65704		

Table 47 - F-Ratio test – Coleman Liau Index (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_NO_OF_RECMNDTIONS

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	3.19193	0.01	0.9257
Denominator	2866	366.49371		

Table 48 - F-Ratio test – SMOG Index (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_NO_OF_RECMNDTIONS

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	1004.11900	2.74	0.0979
Denominator	2866	366.39917		

Table 49 - F-Ratio test – Automated Readability Index (IAD)

Effect of Financial Fundamentals on Total Recommendations

The REG Procedure
Model: INTERACTION

Test INT_EFFECT Results for Dependent Variable r_NO_OF_RECMNDTIONS

Source	*DF*	*Mean Square*	*F Value*	*Pr > F*
Numerator	1	0.03145	0.00	0.9926
Denominator	2866	366.49632		

12 Appendix 4 – Effect plots – (IAD)

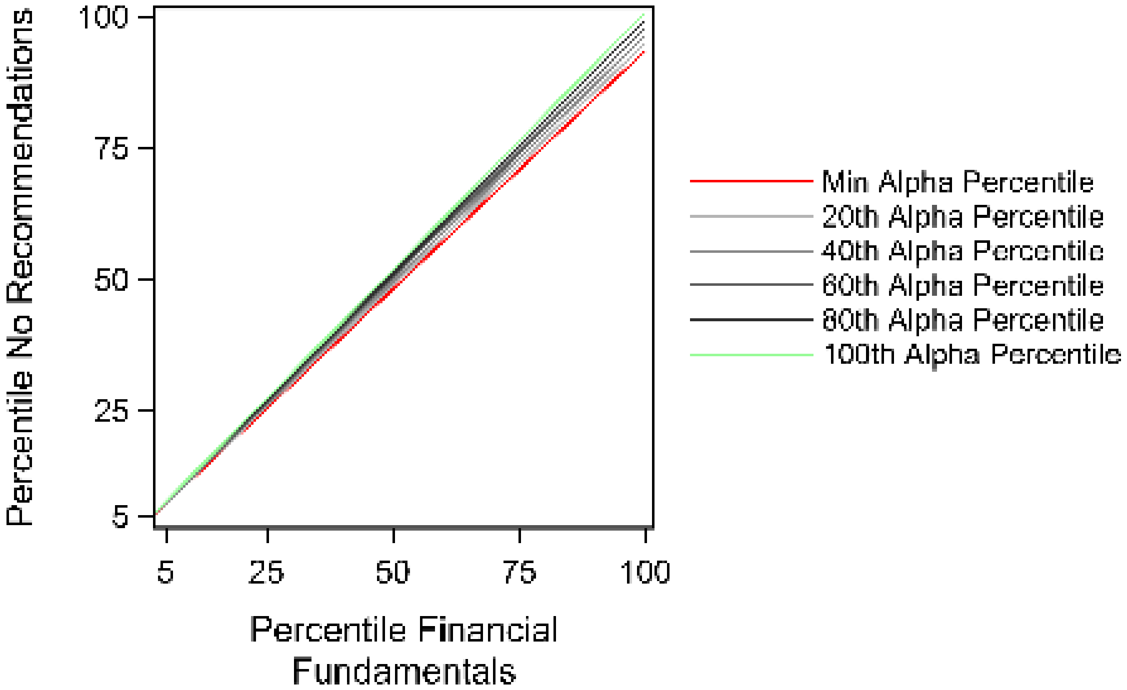


Figure 46 - Effect plot – Alphabeticity (IAD)

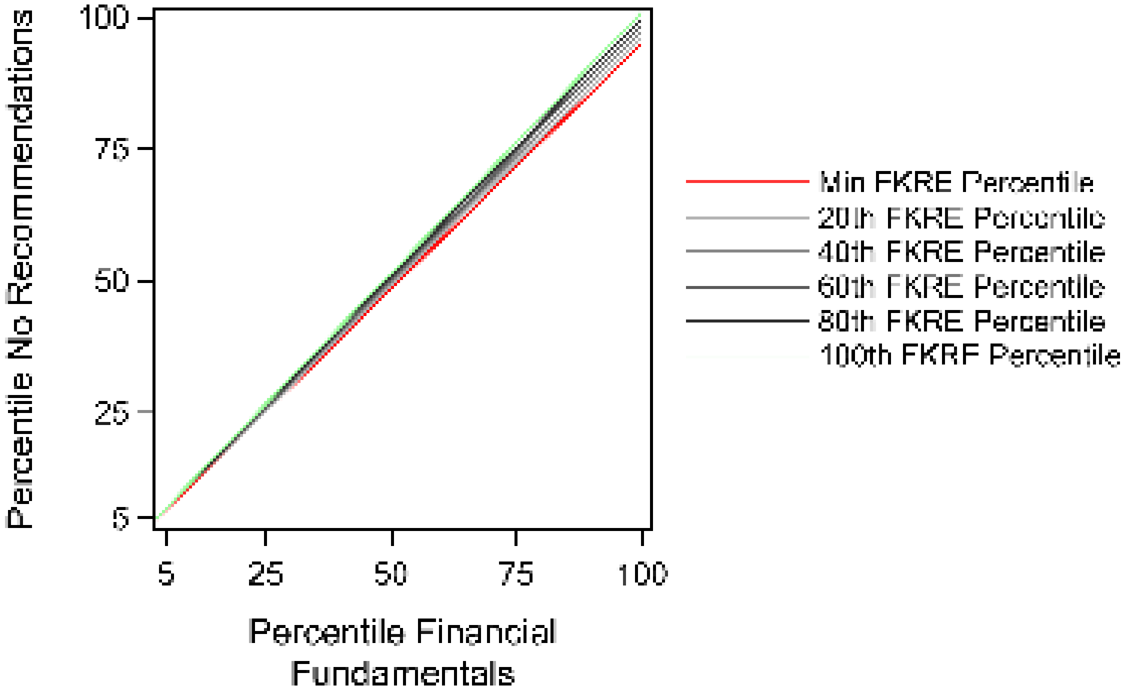


Figure 47 - Effect plot – Flesch Kincaid Reading Ease (IAD)

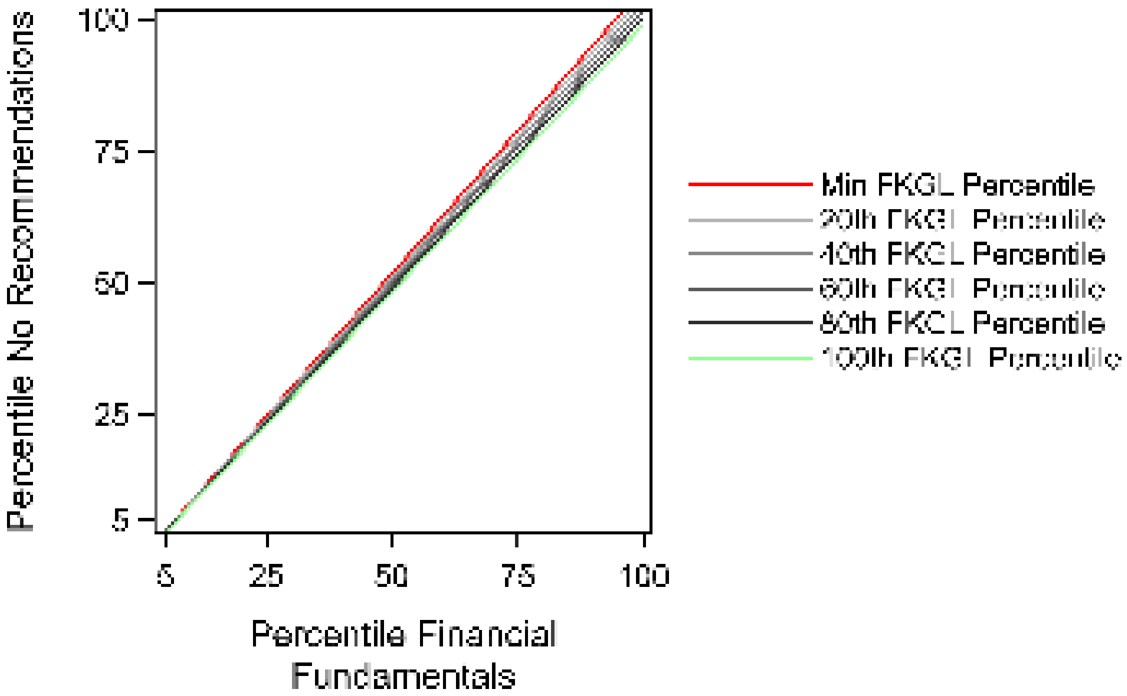


Figure 48 - Effect plot – Flesch Kincaid Grade Level (IAD)

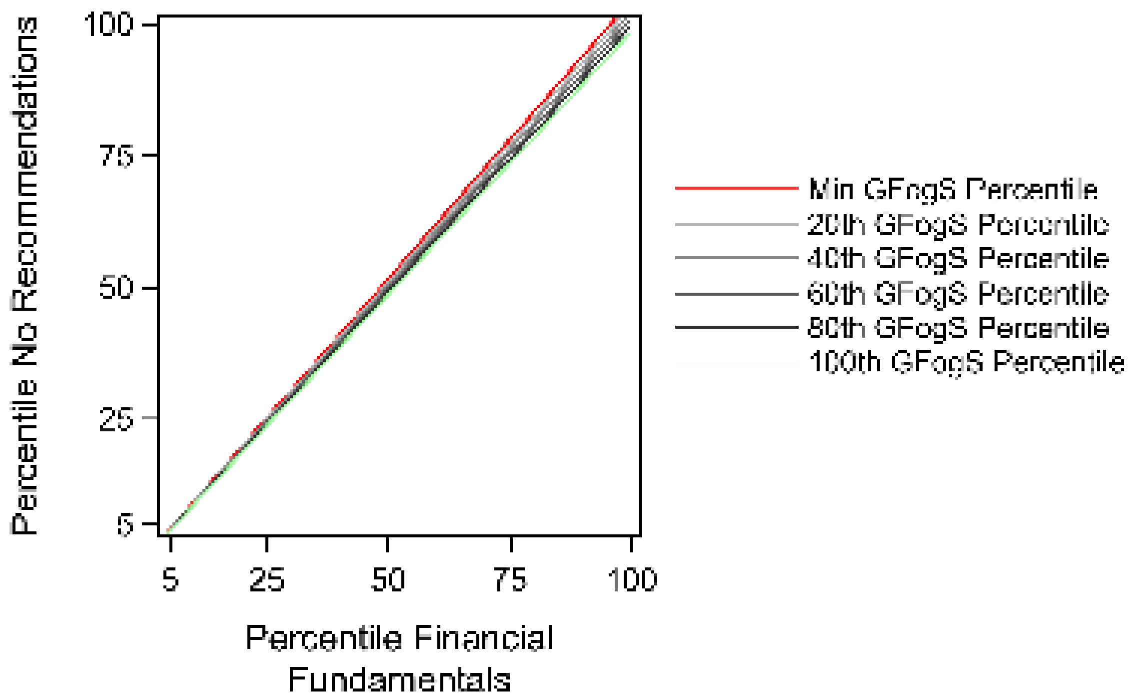


Figure 49 - Effect plot – Gunning Fog Score (IAD)

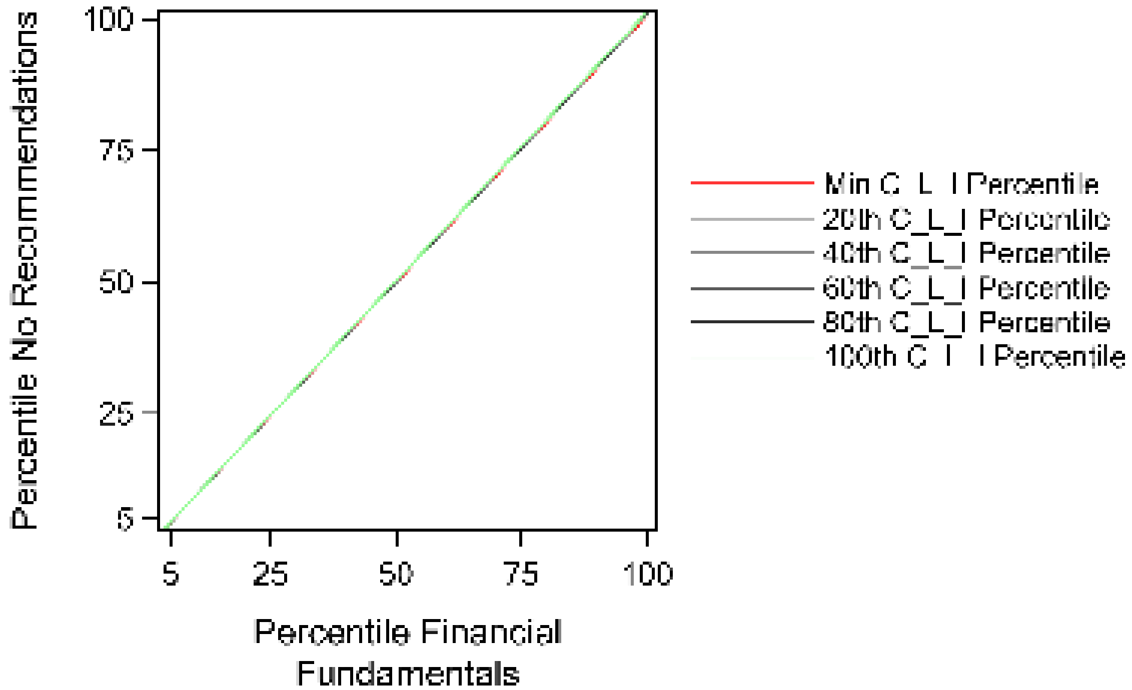


Figure 50 - Effect plot – Coleman Liau Index (IAD)

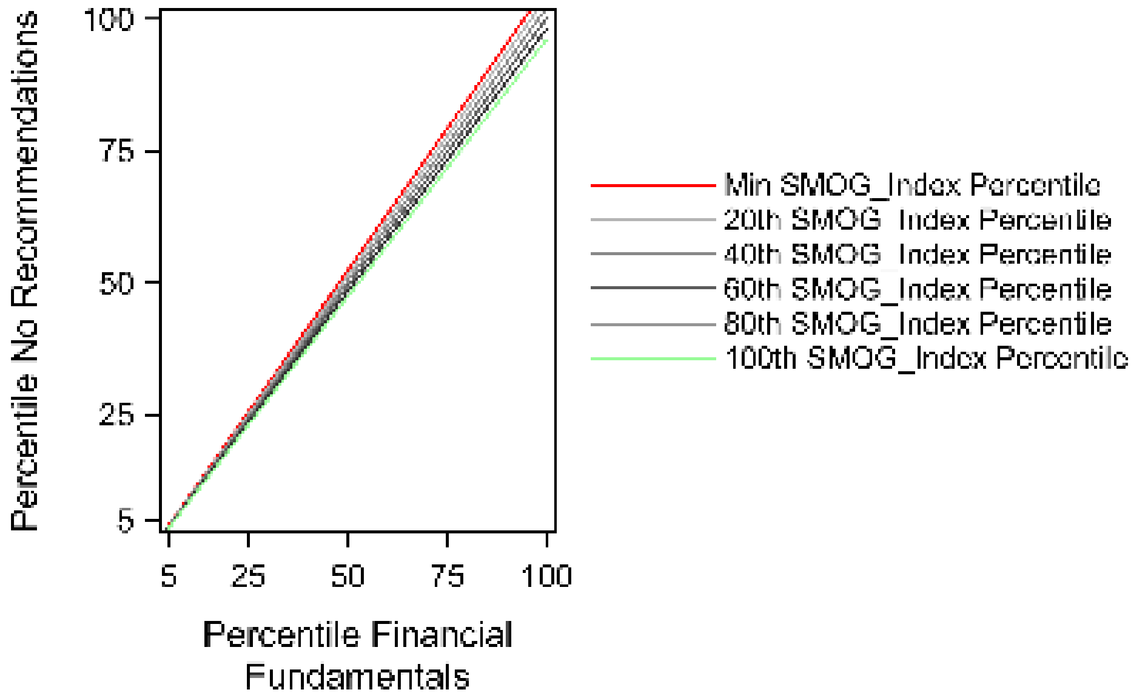


Figure 51 - Effect plot – SMOG Index (IAD)

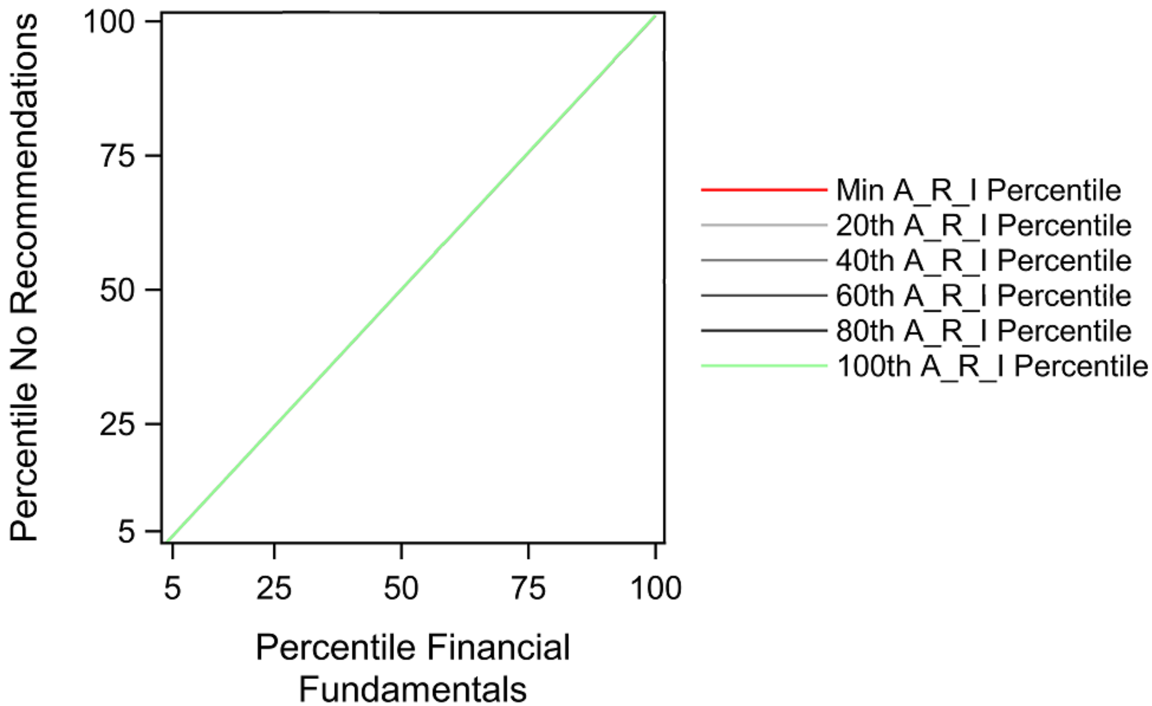


Figure 52 - Effect plot – Automated Readability Index (IAD)

Appendix 5 – Effect plots (Social media data)

12.1.1 Effect plot - (Social Media Data)

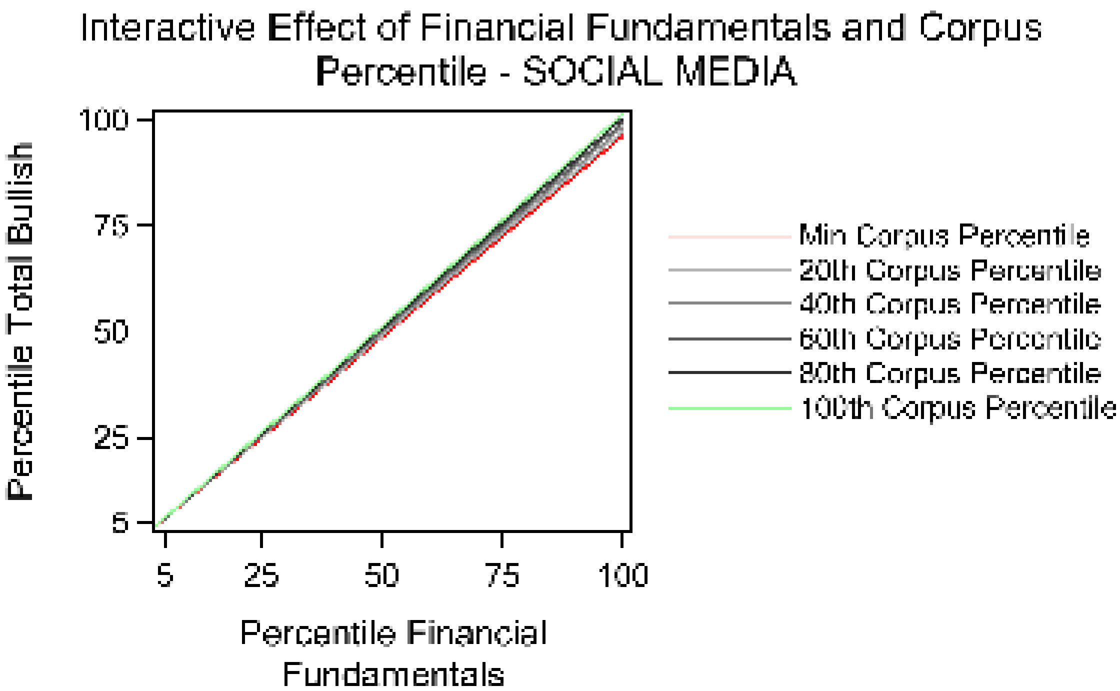


Figure 53 - Effect plot – Corpus Percentile (SMD)

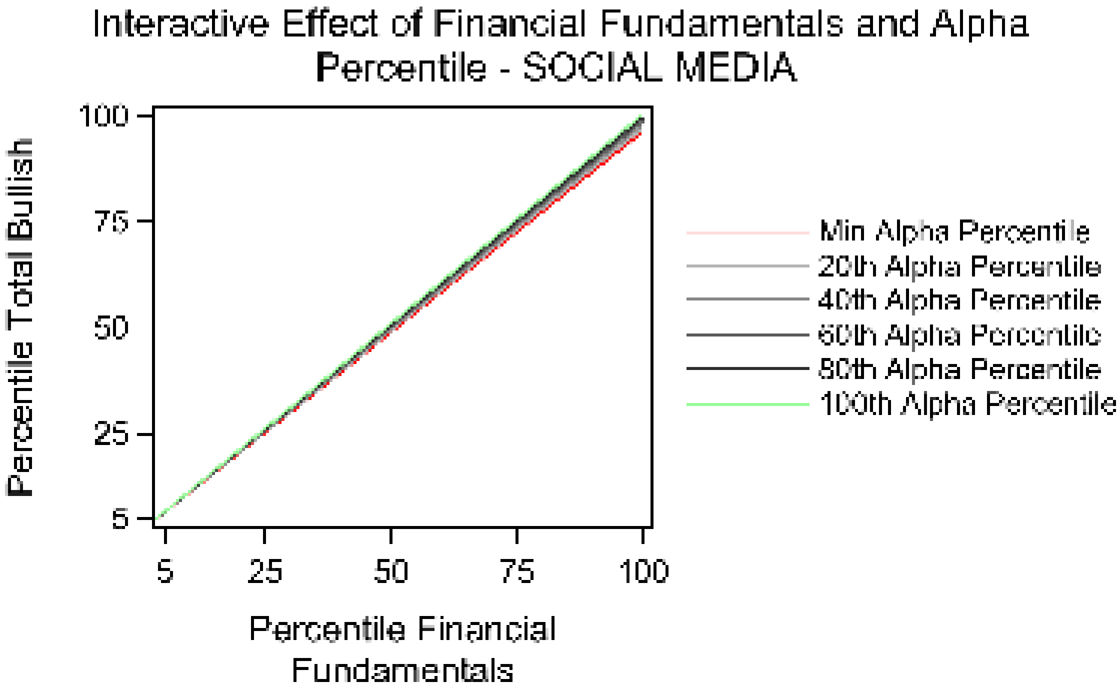


Figure 54 - Effect plot – Alphabeticity (SMD)

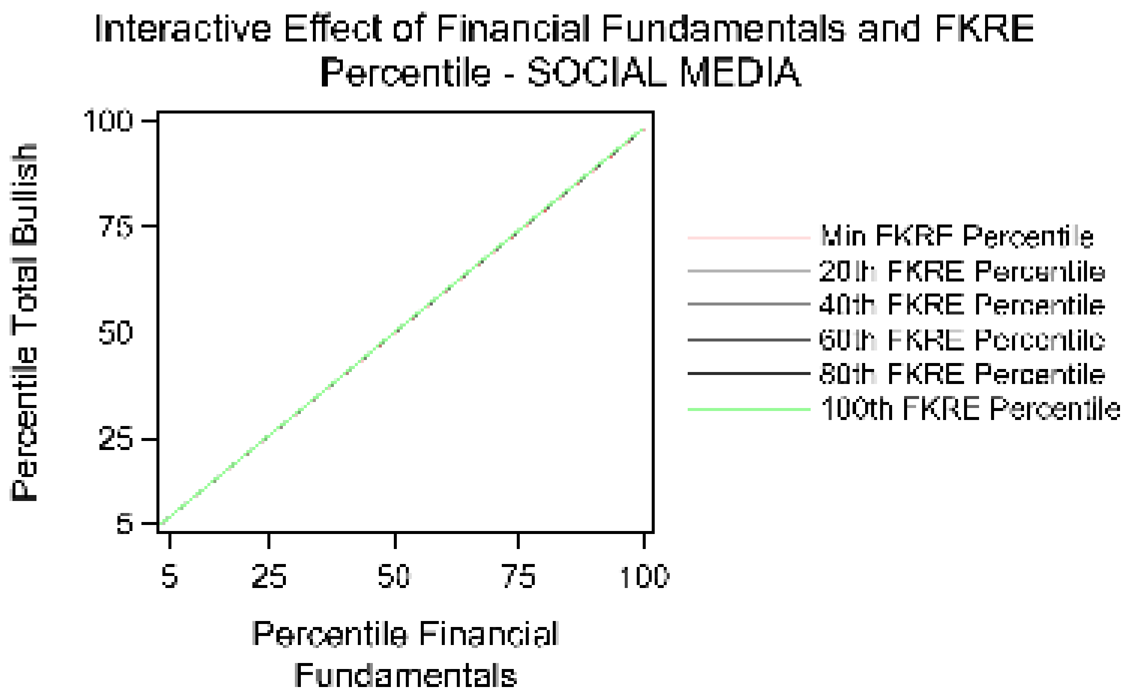


Figure 55 - Effect plot – Flesch Kincaid Reading Ease (SMD)

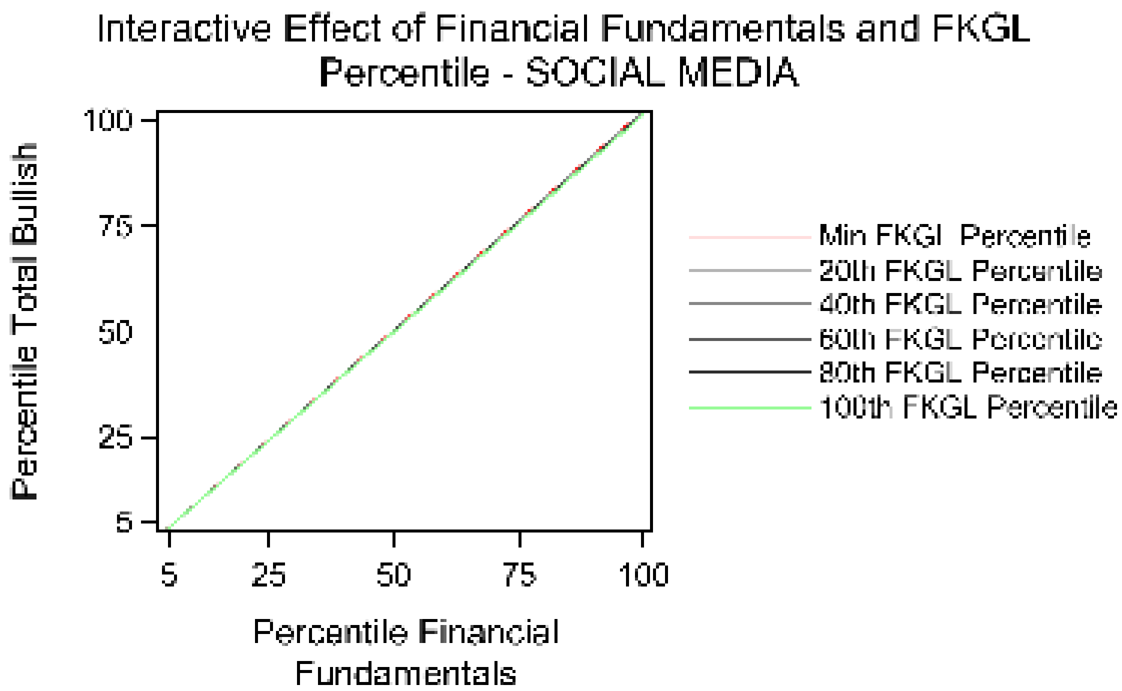


Figure 56 - Effect plot – Flesch Kincaid Grade Level (SMD)

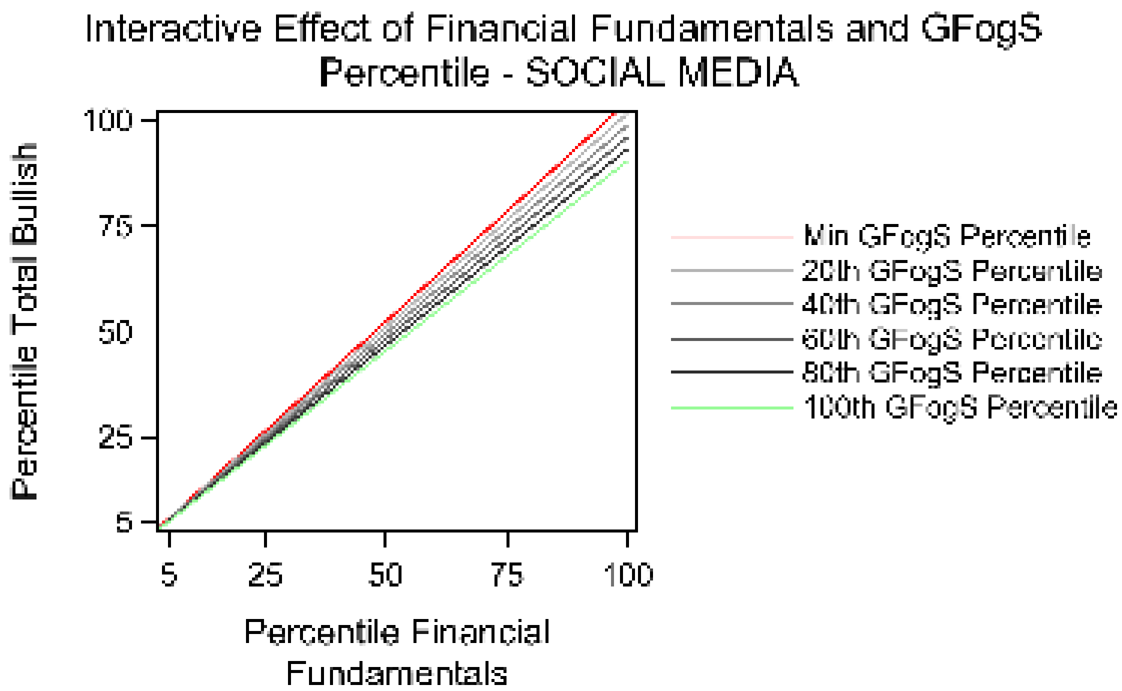


Figure 57 - Effect plot – Gunning Fog Score (SMD)

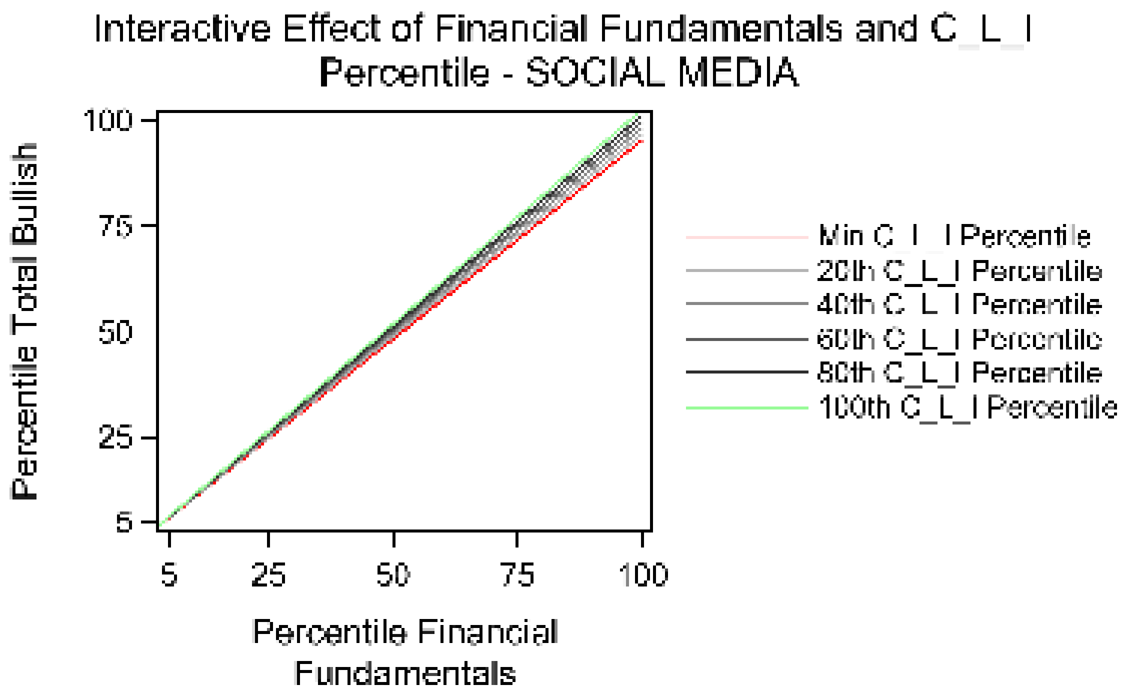


Figure 58 - Effect plot – Coleman Liau Index (SMD)

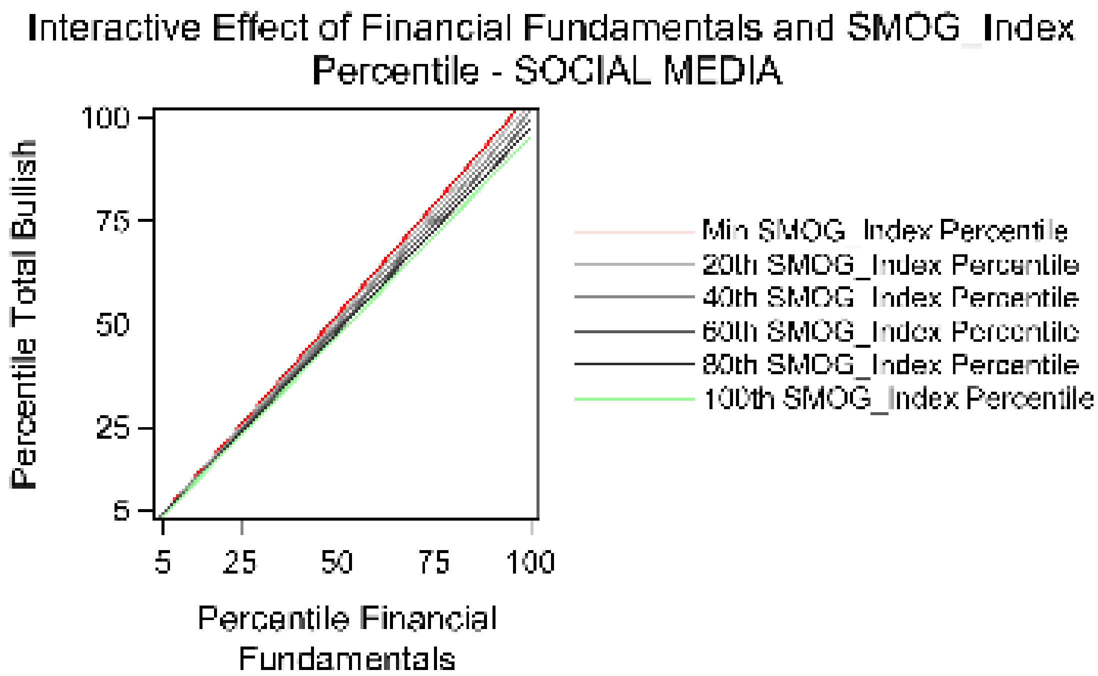


Figure 59 - Effect plot – SMOG Index (SMD)

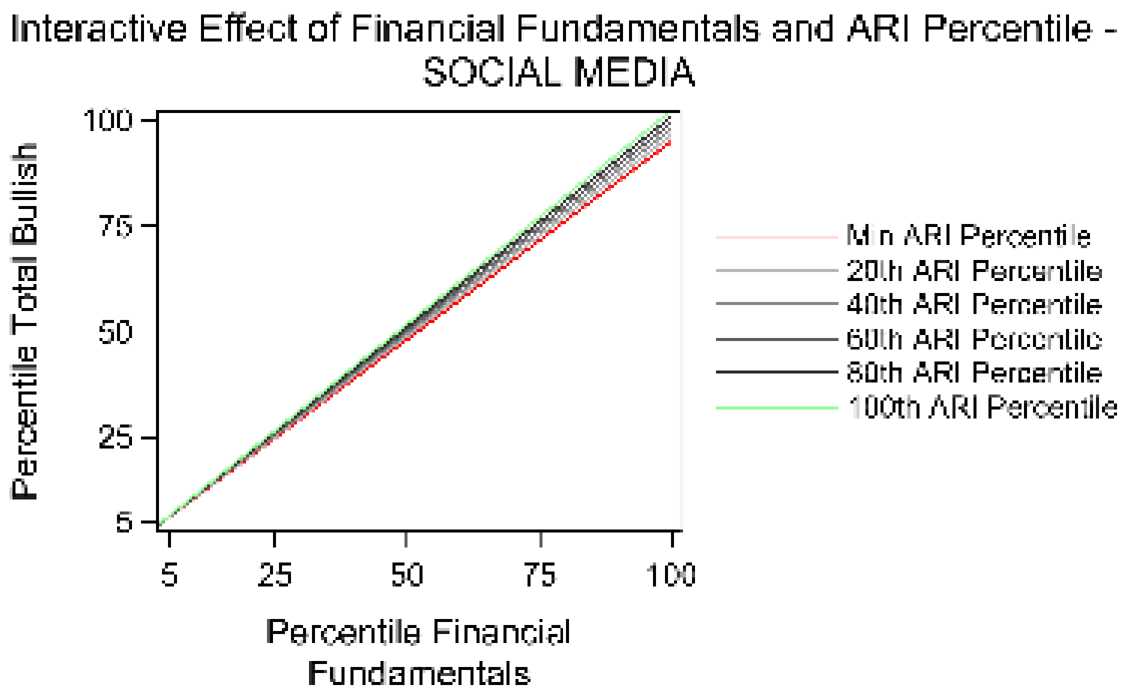


Figure 60 - Effect plot – Automated Readability Index (SMD)

13 SAS Code used to build models

Please refer to CD Rom for the SAS Code used in this research.