Ambiguous Text

Eric Tham*

1 Dec 2019

Abstract

Text is inherently ambiguous. Yet investors read textual news as the primary source of financial information from the financial news and social media. I used Natural Language Processing on social and financial media text to construct a natural event and Big Data ambiguity measurement. The ambiguity measurement is derived from a mixture of distributions model that distinguishes from disagreement between the two sources. A binomial model based on smooth ambiguity preferences is then proposed that explains salient points of ambiguity on asset pricing in empirical tests in this paper and in Brenner and Izhakian (2018). The paper finds that the financial news media have a bigger influence on asset prices than social media except during the last recession from Jun 2009 to Nov 2016. The paper provides a market-wide and natural event evidence of agents' maxmin utility optimisation behavior in Gilboa and Schmeidler (1989).

Keywords: Social media, mixture of distributions, natural language processing, ambiguity aversion, disagreement

JEL Code: D10, D12, G40, G41

*eric.tham@edhec.com, Edhec Business School & National University of Singapore. The author will like to thank Laurent Calvet and Abraham Loiui for their guidance and encouragement, and also to Raman Uppal and Kim Peijnenburg for their helpful comments. He also thanks Richard L. Peterson for access to Thomson Reuters data. All errors are solely mine.

1 Introduction

Technology has impacted how investors receive information and make decisions in the financial markets. This information can be categorised into tangible and intangible news as in Daniel and Titman (2006). Both impact the markets. The news media is a propagator of largely intangible news which is information aside from the quantifiable macro-economic and company fundamentals. Intangible news includes public news and rumours about the company in the form of text. The use of Internet and social media especially amongst the retail investors has particularly democratised the dissemination of information and also increased its speed of transmission. This is especially true for the avid millennial and Gen Z generation users. This raises key questions - how do investors process such information in the midst of diverse information sources and frequently ambiguous textual news?

This is especially important since the popular use of social media from the late 1990s. An increasing number of Americans are turning to the social media as a source of information.[1] Social media seeks to level the playing field for the household investors by improving financial literacy and lowering the costs of household participation. This helps to enhance their social welfare by participation in the stock markets. Yet notably the Internet and the social media is a 'free for all' where views expressed could be misleading or fake. Its information quality can be ambiguous and suspect.[2]

This paper proposes a Natural event and Big Data ambiguity[3] measurement from Natural Language Processing[4] of millions of Internet posts. These posts provide a natural platform for the opinion scores reflecting the views and beliefs of investors. I use a

[1] This Pew centre article shows that social media use amongst Americans increased from 5% of the population in 2005 to almost 70% in 2018.

[2] About two-thirds of Americans get news on social media but most do not trust the news there as accurate - Pew article on 18 Sep 2018.

[3] In this paper, ambiguity and uncertainty are used interchangeably as referring to Knight's classic work in the 1930s while risk follows the Keynesian definition much described in literature.

[4] Natural Language Processing is the computational quantification of large amounts of textual data into a structured format for analysis.

mixture of distributions model to aggregate the opinion scores from two primary information sources - the financial news and social media texts on the S&P 500 index to build an ambiguity measurement. This measurement derives the first moment by a mixing parameter weighted average of the opinion scores and the second moment as the ambiguity itself. This ambiguity is distinct from the disagreement between the two opinion scores although they are positively related. Regression tests are then performed on these ambiguity measurements to quantify which of the information sources is more important. My results indicate the greater importance of the financial news media source since Jan 1998. My results also report broad empirical evidence of the investors' maxmin behaviour that is cited in Gilboa and Schmeidler (1989) that reflects conservative beliefs under ambiguity. This contributed to the greater importance of the social media post the Lehman crisis from Jun 2009 to Nov 2016 due to its more pessimistic tone.

I further develop a simple binomial tree model of ambiguity that explains the salient features of ambiguity on asset pricing based on the smooth ambiguity model of Kilbanoff et al. (2005). The model predictions are consistent with the empirical findings in this paper and also in Brenner and Izhakian (2018). Notably, it explains investors' love for ambiguity in recessionary times and dislike for ambiguity in favourable times. It also predicts a higher ambiguity premium in directional and volatile markets although this is not necessarily in evenly-keeled markets with equal up and down probability markets.

Other most related papers to this paper are Ghysels et al. (2009) and Epstein and Schneider (2008). The paper similarly exploits the linear relationships amongst volatility risks, ambiguity and asset prices in the earlier Ghysels and Epstein papers in the testing.

The paper is organised as follows. In the following section 2, a literature review on ambiguity is first discussed. In section 3.1, the textual data set is introduced including how the difference in opinions or disagreement is obtained from Natural Language Processing (NLP). Its statistical properties are discussed in section 3.2. The binomial tree model to model an agent's behaviour under ambiguity is described in section 4 and the mixture of

distributions to aggregate opinion scores is in section 4.2. An econometric study is in section 5 with a discussion on the results and related events in the 2017-2019. A conclusion on the paper follows with the paper key contributions.

2 Related Literature

In academic literature, ambiguity amongst investors is expressed generally as uncertainty on either consumption, labour market or information quality (different opinions). Whilst the characteristic risk aversion coefficient refers to agents' dislike over the uncertain payoffs, ambiguity can be characterised as their fear of the *unknown unknowns*, where they cannot perceive the *uncertain probabilities* instead.

Empirical literature in Baillon et al. (2018) has shown that ambiguity is a rich phenomenon that is characterised by two main aspects - the well-known ambiguity aversion and the degree of ambiguity, that is the perceived level of ambiguity. The first ambiguity (attitudes) aversion much like risk aversion is a characteristic of the agent. The second concerns the information quality and content, and captures the agent's insensitivity towards prior likelihood changes. When the perceived level of ambiguity is high (for example due to very poor information quality), the less the decision maker is able to distinguish the blurred likelihoods which are then treated as alike. In non-technical terms, too much information (noise) can be no information at all. Both aspects are instrumental in impacting the financial markets, although this paper studies more at the perceived level of ambiguity.

Notably, there was a lack of empirical data in the papers by Epstein and Schneider (2008) and Veronesi (2000) to measure ambiguity. Papers resort to controlled laboratory experiments like choosing urns in the Ellesberg's paradox to quantify ambiguity. The Brenner and Izhakian (2018) paper built an ambiguity measurement by converting high frequency minute price data to lower frequency ambiguity measurement. It documented

several empirical evidences of ambiguity on asset pricing that are replicated with textual data in this paper.

Another paper argued the variance premium to be largely explained by ambiguity aversion Miao et al. (2019). The variance premium is the difference between the risk-neutral and objective expectations of market return variance and results from ambiguity due to compounding effect of market uncertainty. This variance premium largely correlates positively to recessionary periods, which is a similar result of the ambiguity measurement found in this paper.

Ghysels et al. (2009) uses a beta-weighted variance of forecasts of professional forecasters to measure disagreement and used it to proxy ambiguity directly. Ghysels et al recognised that disagreement does not equal ambiguity except under certain conditions. This paper uses a mixture of distributions for the financial news and social media distributions to distinguish between disagreement and ambiguity. Ghysels et al further used several combinations of linear regressions that include volatility risks and uncertainties in the trade-off against returns. This was for the period 1969 to 2003, when the use of Internet was not yet popular. It found a stronger relationship for the uncertainty returns trade-off than the volatility returns trade-off. This is a similar empirical result with our empirical tests using textual data from 1998 to 2019.

3 Financial News and Social Media

3.1 Text as Data

Text as data has gained popularity in the literature. Unlike tangible information which lends itself to econometric analysis, textual data is quantified through features and classified using machine learning into different bins. Prose unlike quantitative data lends itself to subjective and ambiguous interpretation amongst different readers. For example the Wall Street Journal headline on 28 Feb 2019 reads:

Fed Chief Says U.S. Economy 'Is in a Good Place'.

Jerome Powell said the U.S. economy is doing well, but he highlighted risks to growth that prompted the central bank to signal it is done raising interest rates for now.

The ambiguity from text comes not only from within the prose, but also from mixed reviews and opinions that are published by different authors which are propagated through the social media and financial news media. Recent academic literature had showed both the social media and the financial news media to have important impacts on the equity markets.

In the social media, the *tone* of Twitter posts were found to predate stock price movements in Bollen et al. (2011). The volume of Internet search words from millions of households was used in Da et al. (2015) to construct a sentiment index that was found to predict stock price movements and volatility. Over the last decade, social media has evolved to leverage on budding retail investors' interest. Chen et al. (2014) found that the posts from the popular social media investing site, SeekingAlpha.com predict future stock returns and earnings.

The financial news media has been found to impact the financial markets too. In Tetlock (2007), the proportion of pessimistic words on a popular Wall Street Journal column was found to predict downward pressure on market prices. A more recent article by Heston and Sinha (2017) studied 900,000 news stories and found that its sentiment predicts short-term returns, with positive news increasing stock returns quickly and negative news receiving a long-delayed reaction.

Recent research in Kelly et al. (2017) quantifies text as data through a count of the relevant words that appear in articles. I used textual data from Thomson Reuters MarketPsych indices (TRMI), which is a global standard for textual mining in financial markets. TRMI uses a proprietary reference bible of labelled positive or negative words and

semantic inference rules to quantify textual content for opinion scores. Its methodology is similar to the Loughran and McDonald (2011) which uses the General Inquirer database. TRMI sources are however more extensive coming from 2000 news and 800 social media platforms around the world.[5] The TRMI opinion scores range from -1.0 to 1.0, and is daily aggregated from the different media websites.

Two different types of TRMI indices are used in this study - the financial news media and social media posts. The financial news media sources include for example, the Wall Street Journal, MarketWatch and the Financial Times while the social media sources include Facebook, Twitter, SeekingAlpha and etc. The news media sources are generally more objective and factual reporting on fundamentals regarding the economy and companies. They are *perceived* more trustworthy compared to social media posts.[6]

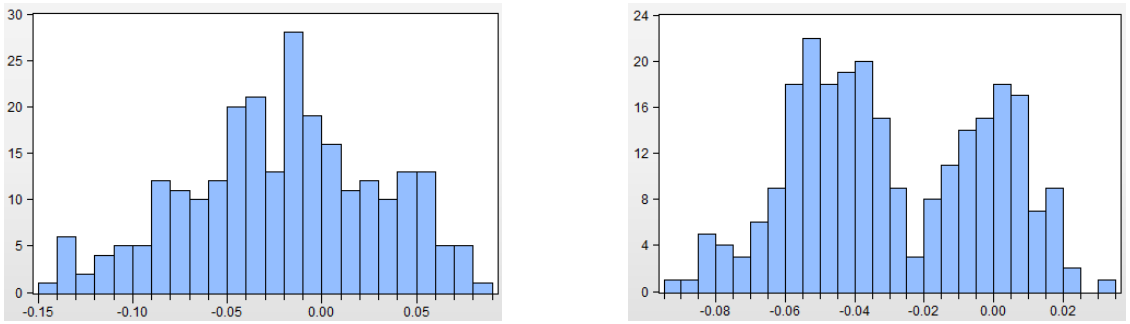
3.2 Statistical properties of text data

The histogram of the monthly opinion scores and the statistical properties of the news and social media opinion scores for the S&P 500 index are in figures 1a and 1b respectively. Notably, the opinion scores in social media relative to news media opinion scores tend to be in *unison* with a smaller standard deviation than the scores from the news media. This hints of a herding effect amongst household investors Nofsinger and Wias (1999). The bimodal distribution of the social media opinion scores indicates the household investors rather post extreme positive or negative comments than neutral comments. The relatively positive skewness indicates a higher tendency to post more positive posts. This contrasts with the relative negative skewness from the financial news media. An interesting article from the Brookings institute writes that bad news sell. Psychology shows that readers

[5]Some literature especially in the computer science domain uses sentiment to mean opinion. To avoid confusion with market sentiment, this paper uses opinion but whether sentiment or opinion, the idea of different opinionation and ambiguity is true.

[6]The journalists from these mainstream news follow a strict code of ethics that strives to ensure the free exchange of information that is accurate, fair and thorough. See the Society of Professional Journalism website.

are drawn to bad news without realising it. Let the suffixes $[m, n]$ denote social media and news media scores henceforth respectively. The monthly correlations from 1998 Jan to 2019 Mar amongst the opinion scores and with the S&P 500 returns are $\rho(s_m, s_n) = 0.25$, $\rho(r_{s\&p}, s_m) = 0.23$ and $\rho(r_{s\&p}, s_n) = 0.21$. A time series of the opinion scores of the social media and financial news media, and the difference in their scores - disagreement is in figure 2. Generally, the social media opinion scores have been more negative than the news media especially since the last recession in 2008.



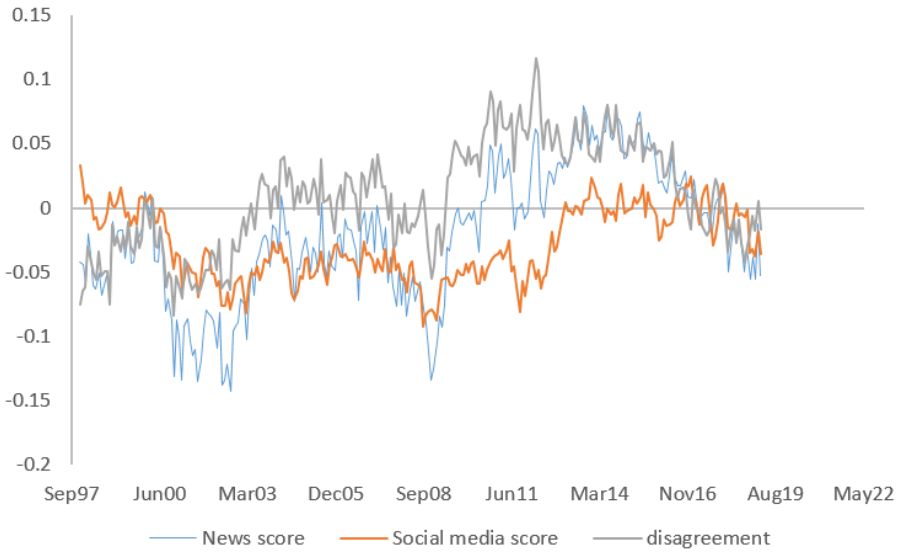
(a) Distribution of news media opinion scores (b) Distribution of social media opinion scores

Figure 1: Histograms of opinion scores from social media and news media

Statistic	News Media	Social Media
Mean	-0.02	-0.03
Median	-0.02	-0.034
Maximum	0.08	0.034
Minimum	-0.14	-0.09
Standard deviation	0.05	0.027
Skewness	-0.17	0.087
Kurtosis	2.47	1.97
Jarque-Bera	4.26	11.4
Probability	0.11	0.003

Table 1: Statistical properties of social media and news media opinion scores

The table shows the statistical properties of the monthly opinion scores on the S&P 500 index by the financial news media and the social media from 1998 Jan to 2019 March. Total number of data points is 255.



There is a general trend of greater negativism in the social media relative to the news media post the Lehman crisis. This corresponds to a larger disagreement between the two information sources as well. In the 2017 to 2019, there is lesser disagreement. A speculative reason is the growing global counter actions and concern against misinformation since 2017. For example the United States in 2017 required the social media giants Facebook, Twitter and Google to testify on their roles in spreading misinformation. The European Union crafted a policy report on *dis*information in 2018.

Figure 2: Time Series of opinion scores and disagreement in the global media on the S&P 500 index from Thomson Reuters MarketPsych

4 Ambiguity Model of Utility

4.1 Binomial Tree model of ambiguity

I next use a binomial tree based model to model a representative agent's behaviour under ambiguity. The model is able to capture the salient features of ambiguity in asset pricing both in this paper and Brenner and Izhakian (2018).

Consider an economy with a single risky asset with an initial price of $P_0 = 1$ at time $t = 0$. Two or more sources (denote as m, n) provide news that impact future cash flows arrive at $t = 0$. There is a representative agent that receives an ambiguous signal s with

smooth ambiguity preferences as in Kilbanoff et al. (2005) of a double exponential form.

$$V(f) = \int^{\Delta} \phi\Big(\int_{s\in\{m,n\}} u(f) d\pi \Big) d\epsilon \tag{1}$$

The f is a real valued function that denotes the agent's plan of action. He chooses to either act on the news from source m or n, or combine a mixture of the sources on the probability measure π that is on S. Each of this news source is ambiguous in nature. Assume the agent has a risk neutral utility function $u(f)$.[7] ϕ is a second order mapping of the probability on the action f. The possible ϕ states are over ϵ between $1-p > \epsilon > -p$ with ϵ as the uncertainty adjustment on the higher payoff U completing the probability space. Consider a binomial case where the cashflows become either U with true probability $p+\epsilon$ or D with probability $1-p-\epsilon$ at $t=1$ for $U > D$. The ϵ is due to the ambiguity of the signal from textual news and different information sources as *perceived* by the agent. The ϵ is similar to the case of posterior parameter uncertainty cited in Gilboa and Schmeidler (1989) and Coles et al. (1995) in that it impacts both the mean and variance of the future cashflows *perceived* by the agent. The preferences function of the agent reduces to:

$$V(f) = \int_{-p}^{1-p} \phi\Big[(p+\epsilon)U + (1-p-\epsilon)D\Big] d\epsilon \tag{2}$$

In the case of ambiguity neutrality, ϕ is linear which will reduce to the standard asset pricing model. For the case of ambiguity aversion, following the ambiguity measurement in Brenner and Izhakian (2018), I use a second order probability functional form as $\mho^2[r] = \int \mathbb{E}[\phi(r)] Var[\phi(r)] df$. This variance of the probability has been used much in literature to express uncertainty for example in the exercise of employee stock options Izhakian and Yermack (2017) and capital structure in Izhakian et al. (2017). The agent expected present

[7] In the case of risk aversion, there would be an additional probability adjustment to risk neutral probabilities. However this would unnecessarily complicate matters with the need for market completeness when the primary point of this model is to illustrate the price adjustment due to ambiguity alone.

utility $U_t(\epsilon)$ ex-post the ambiguious news arrival now becomes in equation 3. This can be represented pictorially in the tree diagram of figure 3.

$$U_t(\epsilon) = \mathbb{E}_t[U * (p+\epsilon)^2 + D * (1-p-\epsilon)^2] \tag{3}$$

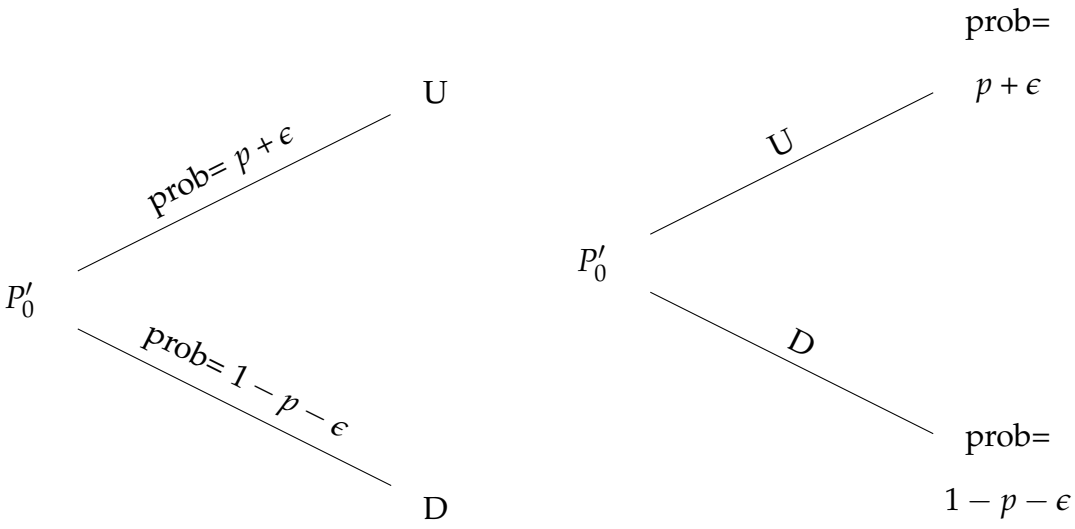


Figure 3: Tree based model of smooth ambiguity

The left diagram in the figure 3 shows a standard asset pricing case (or ambiguity neutral case) where the price P_0' moves up to U with probability $p+\epsilon$ and D with probability $1-p-\epsilon$ at time $t=1$. The right diagram is a consequence of the agent's utility function in equation 3. Instead of weighing the uncertain payoffs by their probabilities, the agent weighs the uncertain probabilities by their payoffs instead. The economic intuition is the agent weighs the importance of their particular beliefs (or probabilities in this context) by its associated payoff. The market is efficient and the asset price changes immediately to the ex-post fair price P_0' after the arrival of news from the information sources. This ex-post fair price P_0' is given by equation 5 which is the expectations of the payoffs in time $t=1$ considering information ambiguity ϵ. There is only one representative agent in the

market for clearing prices.

$$P_0' = (p + \epsilon) * U + (1 - p - \epsilon) * D \tag{4}$$

$$= \underbrace{p * U + (1 - p) * D}_{expected\ drift} + \underbrace{\epsilon * (U - D)}_{ambiguity\ adjustment} \tag{5}$$

This ambiguity adjustment has the useful property of mean-preserving spreads in probabilities analogous to Rothschild-Stiglitz's risk attitude towards mean preserving spreads in outcomes. The agent reacts to the uncertainty by minimising his utility $U_t(\epsilon)$ by solving on his subjective uncertainty ϵ.[8] This is similar to max-min optimisation and is the limiting case of the smooth ambiguity model by Kilbanoff et al. This is done by setting $\frac{\partial U_t(\epsilon)}{\partial \epsilon} = (p + \epsilon) * U - (1 - p - \epsilon) * D = 0$. Solving for ϵ results in equation 6.

$$\epsilon = \frac{1 - p(U + D)}{U + D} \tag{6}$$

The equation 6 expresses the uncertainty ϵ as a function of p. This is an empirical conclusion in Brenner and Izhakian (2018), whence ambiguity is contingent on the expected probability of favourable returns. A higher p reduces the ϵ perceived by the agent. Intuitively, this means with favourable probabilities for higher returns, the agent becomes more ambiguity averse and *allocates* a smaller ϵ to the higher returns scenario. In the opposite case of an unfavourable scenario or 'down' markets, the agents becomes ambiguity loving which results in a higher probability p value and an increased ϵ.

Substituting equation 6 back to the price equation 5.

$$P_0' = \left[p + \frac{1 - p(U + D)}{U + D}\right] U + \left[1 - p - \frac{1 - p(U + D)}{U + D} U\right] * D \tag{7}$$

Further, the derivative $\frac{\partial P_0'}{\partial \epsilon} = U - D > 0$. This positive derivative means that as uncer-

[8] See Gilboa and Marinacci (2013) for an excellent treatise on the history of subjective probabilities and why Bayesian probabilities are not necessarily more rational in an ambiguous setting.

tainty ϵ increases, the price P_0' increases. This reduces future returns, so that expected returns $\frac{\partial r_t}{\partial \epsilon} < 0$. This is consistent with the empirical results later in the paper where the coefficients of the returns to ambiguity measures are all negative. This is a same result in table 4 of Brenner and Izhakian (2018).

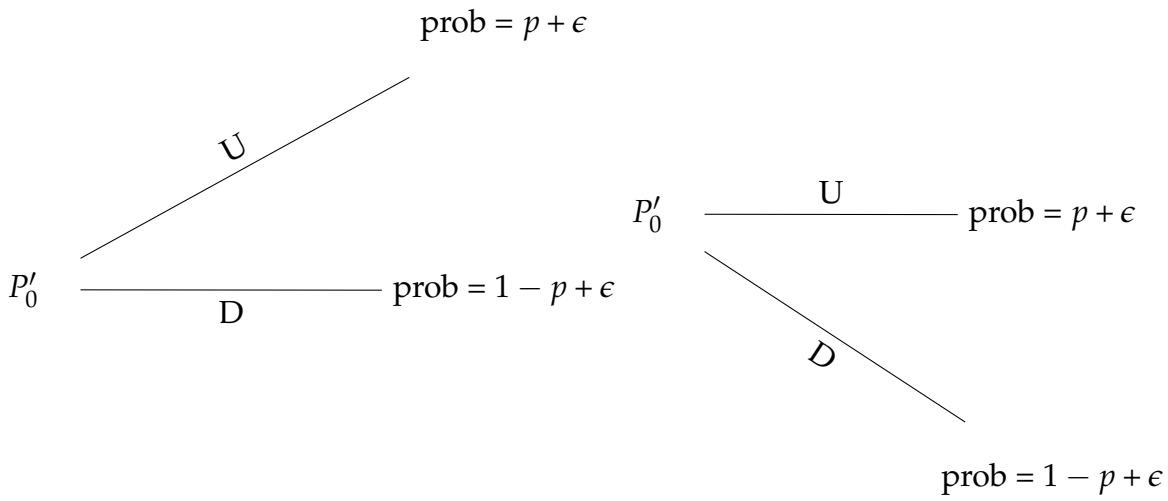


Figure 4: Ambiguity premium $\Delta P_0' < 0$ in upwards and discount $\Delta P_0' > 0$ in downwards markets

Some typical values are applied to the tree model equations 6 and 7 and illustrated in figure 4. Suppose $p = 0.6$, $U = 1.10$ and $D = 1.05$. In this case, $\epsilon = -0.135$ and $P_0' = 1.073$. Without ambiguity, $P_0' = 1.08$ marking an ambiguity discount of 0.007. Note this represents favourable probability since $p > 0.5$ on the higher payoff, and implies the agents' ambiguity averse behaviour (since $\epsilon < 0$).[9] Suppose now instead an unfavourable probability on the higher payoff where $p = 0.4$, $U = 0.95$ and $D = 0.85$. In this case, $\epsilon = 0.155$ and $P_0' = 0.905$. Without ambiguity, $P_0' = 0.89$ marking an ambiguity premium

[9]Whilst it is not the intent of this paper, the results are consistent with the empirical studies linking ambiguity aversion and over-confidence. In this model, over-confidence occurs when the agent places a higher than objective probability p weight on the larger cashflow U in the favourable scenario. Brenner et al. (2015) found in an experimental study that ambiguity has a negative impact on overconfidence. This is evidenced in the model with $\epsilon < 0$ for the favourable cashflow scenario U. Overconfidence has been attributed in literature to two main causes - agents' ability and ambiguity. The modelling of over-confidence and ambiguity in this manner however is flimsy since it is based on a single parameter ϵ.

of 0.015, implying the agent's ambiguity loving behaviour since the agent is willing to pay more with ambiguity.

I further use this model to make predictions. Firstly, when the probability $p = 0.5$, the ambiguity premium becomes less significant, especially if the up and down price movements offset each other. This happens for example when $U = 1.1$ and $D = 0.9$ and the ambiguity premium is actually zero. This makes intuitive sense since any risk the agent perceives in such an evenly keeled market is allocated as the risk premium. This evenly-keeled market is applicable even for large $U - D$, say $U = 1.5$ and $D = 0.5$. This implication means that even in a very volatile market but the odds are even, the ambiguity premium can be still zero. Mathematically, the Intermediate value theorem states for a continuous function $f(x)$ with an interval $[a, b]$ that takes values $f(a)$ and $f(b)$ it must also take any value between $f(a)$ and $f(b)$, which in this particular case implies the ambiguity premium/ discount must take on the value of zero.

On the contrary, the ambiguity premium is large when the expected movement differences $U - D$ are large, and they are moving in the same direction. This happens for example when $U = 1.2$ and $D = 1.1$ or $U = 0.9$ and $D = 0.8$. Naturally this also implies that the ambiguity premium will be greater in directional and more volatile markets (for large $U - D$). There is evidence in Ghysels et al. (2009), Miao et al. (2019) and in figure 8 of this paper when the ambiguity index is shown to be the highest prior and during the recessionary times.

The model offers a natural explanation for the significance of the ambiguity premium during recessionary times and *not* in expansionary times since it depends directly on the $U - D$, which is a proxy for the dispersion or volatility of the returns. This is due to the asymmetrical volatility in markets cited in Bekaert and Wu (2000) which makes the conditional variance of returns correlates negatively with returns in the presence of negative news. When returns decreases in recessionary times, the volatility increases with the ambiguity premium increasing consequentially. On the contrary when returns are high in

expansionary times, volatility does not move as much due to assymmetric volatility with the consequential ambiguity premium being much less. The model further inherits several useful features of the smooth ambiguity model in that it dissects amongst ambiguity aversion (second order probability functional form), perceived information ambiguity (ϵ) and risk aversion (neutrality in this example).

4.2 Mixture of distributions hypothesis for ambiguous sources of information

The previous binomial tree model illustrates how a representative agent behaves under ambiguity. I further use the mixture of distributions model to aggregate different information sources and their signals $s_{i,t}$ the agent receives. The mixture of distributions has similarly been used in determining the empirical density of option prices reflecting the beliefs distribution of different market agents. See Bahra (1997). Consider now the signal at time t for $i \in \{m, n\}$ in equation 8 on the dividend rate d_t:

$$s_{i,t} = d_t + \epsilon_{i,t} \qquad \epsilon_{i,t} \sim N(0, \sigma_i^2) \tag{8}$$

I test both the news media and social media on their signals $s_{i,t}$ individually to determine which is more important in impacting ambiguity. The monthly level of ambiguity is calculated by using the means and standard deviations of the daily signals that are assumed normal. This conversion from high to low frequency approach has been used in Brenner and Izhakian (2018) where 5 minute data is used to construct daily distributions which in turn is used to construct the monthly ambiguity. In this case, the daily distributions are used to derive the moments of the monthly distributions in equation 10 described below.

The agent forms prior beliefs from either s_n or s_m or a mixture of them based on a mixture of distributions from these sources. Ignoring the time subscripts for notational simplicity, the first and second moments for the mixture of k distributions are derived

from Fruhwirth-Schnatter (2006). The ω in equation 10 represents the mixing parameter or weight placed on the signal s_n and $1-\omega$ on the signal s_m and thence determines the relative importance of the two signals in the agent's formulation of beliefs.

$$s_\psi = \sum_{i=1}^{k} \omega_i s_i \tag{9}$$

$$\sigma_\psi^2 = \sum_{i=1}^{k} \omega_i (s_i^2 + \sigma_i^2) - s_\psi^2 \tag{10}$$

In the case of the financial news media and social media for $k = 2$ this is represented in equations 11 and 12 respectively. Similarly, higher moments like the skewness and kurtosis for the ambiguity can be computed but are ignored in this study. The second moment σ_ψ represents the perceived level of ambiguity from the information sources.

$$s_\psi = \omega s_n + (1-\omega) s_m \qquad first\ moment \tag{11}$$

$$\sigma_\psi^2 = \omega(s_n^2 + \sigma_n^2) + (1-\omega)(s_m^2 + \sigma_m^2) - s_\psi^2 \qquad second\ moment \tag{12}$$

The use of the mixture of distribution as compared to a convoluted bivariate distributions is appropriate since a convoluted distribution would have reduced the dispersion of the probabilities distribution (and thence ambiguity). Empirically, this would not be correct since with a greater number of competing information sources, the greater the ambiguity would be. Let Φ denote the difference between the two opinion scores in equation 13. This represents the disagreement between the two sources.

$$\Phi_t = s_{m,t} - s_{n,t} \tag{13}$$

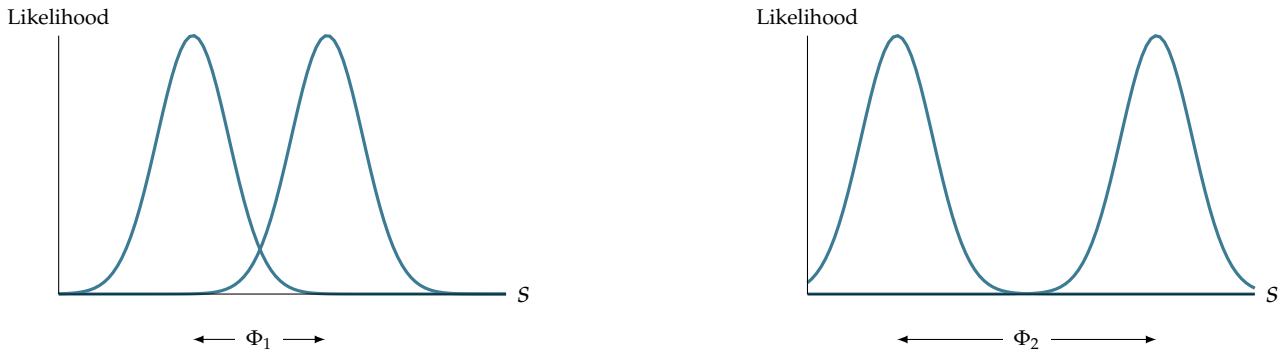


Figure 5: Greater disagreement $\Phi_2 > \Phi_1$ increases ambiguity

The figures show the distributions of two information signals which means are very different. Φ represents the disagreement with $\Phi_2 > \Phi_1$ such that the dispersion in the mixed distribution on the right hand side is larger with greater ambiguity.

Rationally, a greater disagreement between the two signals would result in greater ambiguity. Suppose $\omega = 0.5$ such that the two signals are equally weighted. Thence, $s_\psi = 0.5(s_m + s_n)$ and the second moment in equation 12 would reduce to equation 14 where the cross term $0.25 s_{m,t} s_{n,t}$ is negligible.

$$\begin{aligned} \sigma_\psi^2 =& 0.25\Phi^2 + 0.25 s_{m,t} s_{n,t} + 0.5(\sigma_n^2 + \sigma_m^2) \\ \approx& 0.25\Phi^2 + 0.5(\sigma_n^2 + \sigma_m^2) \end{aligned} \tag{14}$$

An increase in disagreement Φ thence increases the level of ambiguity. Further, information sources with 'large' uncertainty $\sigma_{\phi,i}$ increase the ambiguity, especially if it has greater mixing weight ω_i. Intuitively, this can happen if a rumour spreads widely amongst the community, the agent 'believes' it with a greater *weight* ω_i on it for decision-making. The ω could also be optimised by the agent endogenously to reduce decision making uncertainty or ambiguity much like in the earlier binomial tree model.

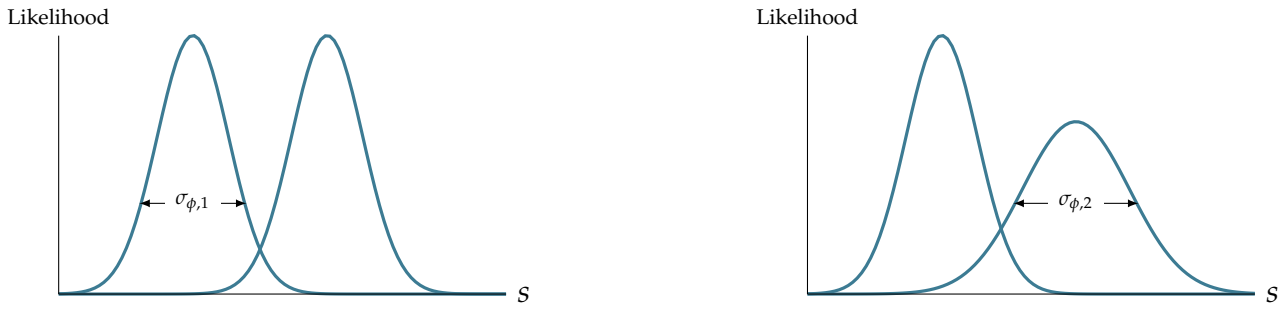


Figure 6: Larger individual information uncertainty $\sigma_{\phi,i}$ increases ambiguity

The figures show the distributions of two information signals with different standard deviations ($\sigma_{\phi,2} > \sigma_{\phi,1}$) such that the ambiguity on the right hand side would also be greater.

Using these measures, an index of the ambiguity scores for the financial news media, social media and from their mixture disagreement are constructed as time series in figure 7 below. The ambiguity from the mixture disagreement sources and the financial news media are notably higher than the social media. This can be attributed to the tendency of social media posts being worded less ambiguous (smaller standard deviation of opinion scores as noted in section 3.2.

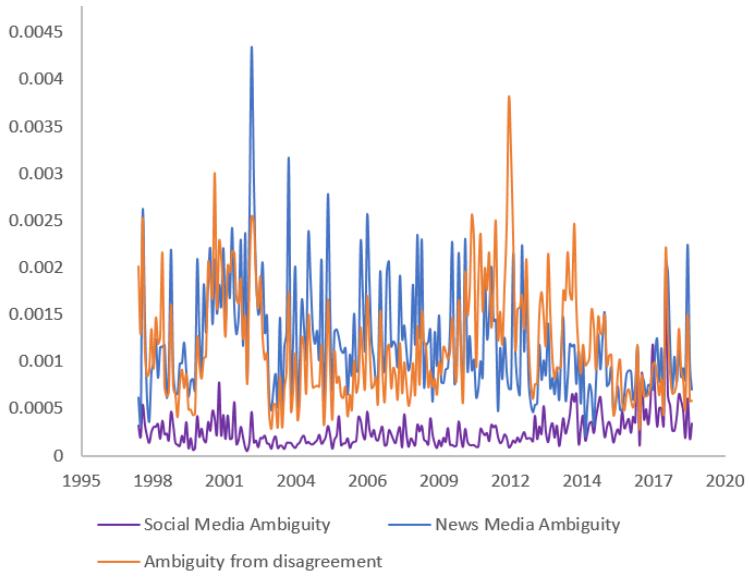


Figure 7: Time Series of ambiguity index from global media on the S&P 500 index

5 Econometric Study

In both Epstein and Schneider (2008) and Ghysels et al. (2009), the returns is derived as a function of the volatility and the degree of ambiguity σ_ϕ in equation 15.

$$p_0(s) = \mathbb{E}[d|s] = m + \gamma_\psi s_\psi + \rho\sigma_d + \theta\sigma_\psi \tag{15}$$

This is similar to the equation 5, with the expected drift term as $m + \gamma_\psi s_\psi$ and the ambiguity premium as the $\theta\sigma_\psi$. Define $\gamma_i = \frac{cov(s_i, r_i)}{var(s_i)}$ as the regression coefficient of the return r against the individual signal s_i. Thus γ_i represents the useful signal to noise ratio of the signal s_i and represents the reliability of the signal. In the empirical section 5.1, it represents the coefficient of the information signal $s_i \; \forall \; i \in \{m, n\}$. A high value represents the usefulness of the signal but not necessarily the ambiguity of the signal, which would be represented by its individual $\sigma_{\phi,i}$.[10]

I test various hypothetical scenarios by OLS regressions on the S&P 500 returns using combinations of variables. The regression results are in table 2.[11]

5.1 Empirical discussion

The model numbers 1 to 3 in table 2 test hypotheses whence the agent observes the opinion score signal individually. The individual opinion score from the financial news s_n has a marginally higher explanatory power than the social media score s_m. A latest survey by Pew center in December 2018 showed that social media still falls behind news websites as a source of information. There is evidence that the agent considers both signals as s_{nm}

[10] A useful illustration is that of an obtrusive lie. The obtrusiveness of it makes it less ambiguous but as a lie it is simply unreliable.

[11] In the empirical study, the GARCH(1,1) is used to model the S&P 500 voatility. The equations for the returns and the conditional variance $\sigma_{d,t}$ are in 16 and 17 respectively.

$$r_t = \epsilon_t \qquad \epsilon_t \sim N(0, \sigma_t) \tag{16}$$

$$\sigma^2_{d,t+1} = c + \theta\sigma^2_{d,t} + \alpha\epsilon^2_t \qquad \textit{Garch equation} \tag{17}$$

which is the equally weighted ($\omega = 0.5$) average of the signals s_n and s_m has the highest log-likelihood. A counter-intuitive observation is the higher γ_m for the social media which indicates it as having a higher signal to noise ratio. This may be attributed to the lower standard deviation of the social media scores as discussed earlier, and also to social media being a propagator of largely stale news in Tetlock (2011). This means investors react to repeated social media information as though they are new. There is similar evidence in Jiao et al. which found that social media acts as 'echo chambers' of existing information.

Model numbers 4 to 6 add the individual ambiguity $\sigma_{\psi,i}$ to the opinion scores for the regression. The inclusion of the ambiguity measurements in all cases provide a better fit. All coefficients of the ambiguity measurements are negative as predicted by the tree-based model. The most probable hypothesis is model 6 with the equal average of both the social media and financial news media scores.

In Model number 7, the returns are regressed against the S&P volatility, σ_d. The coefficient is found to be statistically insignificant - the same puzzling relation between intertemporal risk and return cited in Scruggs (2002) and Campbell (1999). This was attributed in Veronesi (2000) to the opposite effects of ambiguous information. On one hand, a negative news signal that increases volatility depresses casflow expectations but it also increase hedging demand due to agents' risk aversion over the poor information quality. The aggregate impact on the returns from the increase in the volaility is therefore insignificant and even negative. Model number 8 considers this ambiguity-returns trade-off instead of the risk-returns trade-off. The ambiguity-returns is more significant substantiating Ghysels et al's finding although in both Model 7 and 8, the coefficients are not statistically significant.

However when the σ_d is regressed with the signals from the financial news, social media and both the media in model no 9, 10 and 11 respectively, the coefficients becomes positive and significant - a similar finding with Brenner and Izhakian (2018).

The Models 12 to 14 test all three variables simultaneously to ascertain which has the best fit. The most probable model is 14 with the ambiguity from the disagreement between the social media and financial news media scores and the average of the opinion scores. Model 15 with the agent choosing s_{nsm} as the minimum of the two opinion scores shows the best fit model, indicating there is some form of agent optimising behaviour towards ambiguity. It is important to realise that the mixture of distributions in itself does not offer an explanation for how the agent optimises his utility with ambiguity (aversion). Instead, it offers an avenue for the agent to weigh and aggregate his beliefs by choosing ω to reduce his perceived uncertainty.

5.2 Empirical results

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
s_n	0.21			0.17					0.30			0.26			
	(3.42)			(2.64)					(4.66)			(3.84)			
s_m		0.354			0.419					0.450			0.502		
		(2.82)			(3.19)					(3.58)			(3.82)		
s_{nm}			0.323			0.319					0.470			0.489	
			(3.48)			(3.58)					(4.47)			(4.96)	
s_{nsm}															0.509
															4.35
$\sigma_{\psi,n}$				-9.43								-8,40			-4.72
				(-1,88)								(-1.73)			(-0.92)
$\sigma_{\psi,m}$					-31.31								-28.4		
					(-2.14)								(-1.81)		
$\sigma_{\psi,nm}$						-8.17		-8.72						-12.0	
						(-1.81)		(-1.57)						(-2.79)	
σ_d							0.129		0.623	0.457	0.692	0.598	0.423	0.811	0.854
							(0.45)		(2.34)	(1.858)	(2.64)	(2.26)	(1.70)	3.07	(2.96)
c	0.01	0.02	0.01	0.02	0.03	0.02	0.00	0.02	-0.02	0.00	-0.01	-0.01	0.01	0.00	0.00
	(4.22)	(5.09)	(5.38)	(3.50)	(4.67)	(4.29)	(-0.01)	(2.48)	(-1.42)	(-0.14)	(-1.22)	(-0.41)	(0.73)	(-0.36	(-0.36)
Adjusted R^2	0.052	0.045	0.062	0.061	0.057	0.069	-0.002	0.009	0.081	0.061	0.098	0.088	0.070	0.118	0.119
Log-likelihood	438.3	437.4	439.7	440.1	439.6	441.2	429.1	432.7	442.9	440.0	445.2	444.3	441.9	**448.6**	**448.7**

The dependent variable is the monthly excess S&P 500 returns. The terms in brackets are the t-stats that are Newey-West adjusted. Most coefficient are significant at the 5% confidence level. The log-likelihood values in bold for models no 14 and 15 refer to the best hypothesis model. The regression period is for the period 1998 Jan to 2019 March for 255 monthly data points.

Table 2: OLS Regression hypothesis results

5.3 Agent's behaviour under ambiguity

The result from the last section regression - s_{nsm} for Model no 15 hints of the agent reacting towards ambiguity by being more conservative similar to Gilboa and Schmeidler (1989) maxmin utility behaviour. To test this behaviour, two specification regressions are further done in table 3.

In the first specification, the Model numbers 1 to 7 vary the ω from 0 to 1. The Model number 5 and 6 with $\omega \approx 0.55$ shows the highest likelihood fit, highlighting the relative importance of the financial news media to the social media in impacting asset prices.

The last column 'Min' is similar to s_{msn} but dynamically adjusts the weight $\omega_{i,t} = 1$ to correspond to the minimum (most pessimistic) opinion score signal $s_{i,t}$ at each month.

Model No	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ω mixing parameter	0.0	0.25	0.5	0.55	0.75	0.80	1.0	Min
s_{ns}	0.502	0.549	0.489	0.468	0.368	0.341	0.256	0.502
	(3.82)	(4.54)	(4.96)	(5)	(4.90)	(4.77)	(3.84)	(4.75)
σ_ψ	-28.42	-14.49	-12.03	-12.04	-12.42	-12.37	-8.40	-5.27
	(-1.81)	(-2.21)	(-2.79)	(-2.87)	(-2.87)	(-2.76)	(-1.73)	(-1.66)
σ_d	0.423	0.722	0.809	0.81	0.757	0.730	0.598	0.903
	(1.70)	(2.82)	(3.07)	(3.06)	(2.89)	(2.80)	(2.25)	(3.32)
c	0.0089	0.00072	-0.0036	-0.0037	-0.0031	-0.0027	-0.0052	-0.0089
	(0.731)	(0.071)	(-0.362)	(-0.38)	(-0.294)	(-0.257)	(-0.414)	(-0.875)
Adjusted R^2	0.070	0.10	0.118	0.118	0.112	0.109	0.088	0.12
Log-likelihood	441.8	446.8	**448.62**	**448.64**	447.7	447.3	444.3	**448.8**

The dependent variable is the monthly excess S&P 500 returns. The terms in brackets are the t-stats that are Newey-West adjusted. Most coefficients are significant at the 5% confidence level.

Table 3: Testing relative importance of financial to social media news

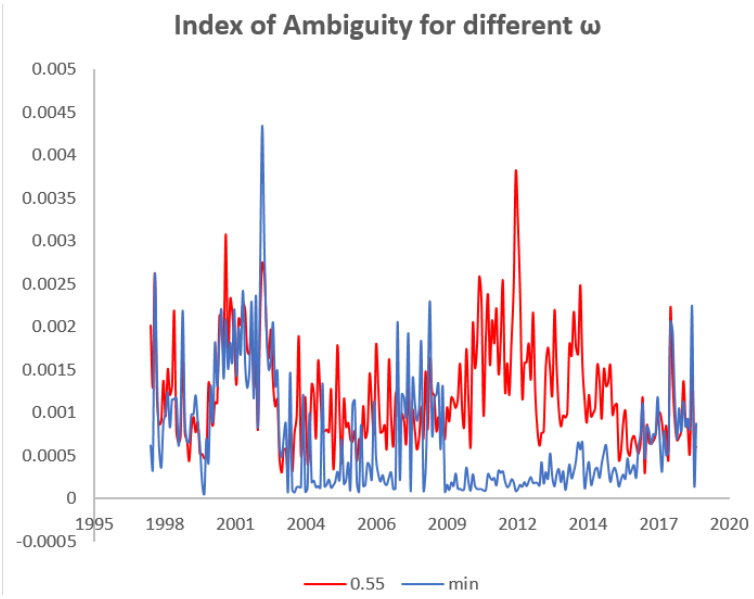
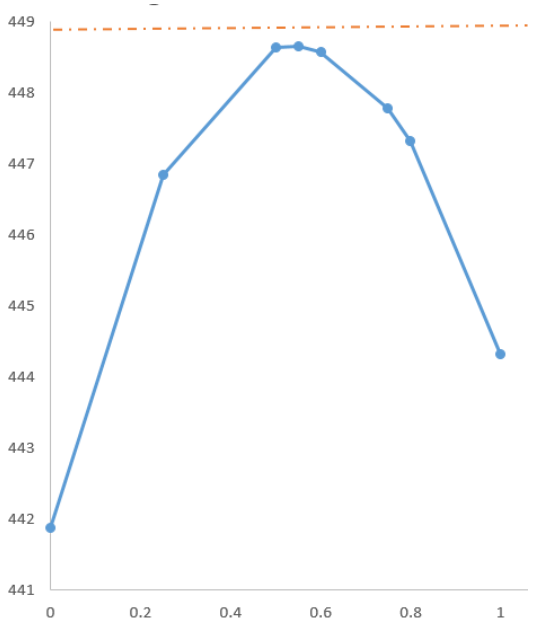


Figure 8: Different ambiguity index based on ω



The dotted line is for the dynamically weighted ω with the most pessimistic belief $s_{i,t}$ in a maxmin-like utility optimisation.

Figure 9: Likelihood values for different mixing parameters ω

The figure 8 shows a time series of different ambiguity indices with $\omega = 0.55$ with

the most optimal likelihood and for the dynamically weighted ω. Interestingly, both the ambiguity indices concur with the ambiguity index cited in Izahakian et al in a Dec 2017 WSJ article. In the article, the ambiguity index constructed from market data hit historical highs in 2017 in spite of the (CBOE) volatility index being at contradictorily low levels. An examination of the figure 2 shows the ambiguity is caused more due to the uncertainty of both the social media and financial news media sources and less by the disagreement of these two sources. Notwithstanding the high ambiguity index, this did not necessarily translate to a high premium if the market odds are evenly-keeled. In such markets, investors exhibit portfolio inertia not knowing what to do either way in the face of uncertainty. This is consistent with the conclusion in Illeditsch (2011) and Ameriks and Zeldes (2000) which documented that investors do not trade much on surprising news.

The figure 9 shows the log-likelihood being plotted against the different ω. The red line is for the dynamically weighted strategy showing it with the highest log-likelihood. This suggests that the agent is more likely to follow the most pessimistic opinion score with its associated ambiguity. Observing the figure 2 for the time series of the different media opinion scores shows that there are long periods that the social media score is more pessimistic than the financial news score from Jun 2009 to Nov 2016. During this period, the agent would be following the social opinion scores with $\omega_m = 1$. This behaviour is especially so during this Lehman crisis and Greek debt crisis but less during the earlier dot-com bust in the early 2000s. This observation suggests the increased importance of the social media in influencing asset prices news during recent economic crises. However, notwithstanding the higher $\omega_n \approx 0.55$ on the financial news media indicates its greater importance in influencing asset prices. It is likely that there are agents in the market that follow both strategies.

The relative importance of the financial news media becomes greater however when weekend news is used to construct the ambiguity index[12] with the optimal $\omega \approx 0.7$. The

[12] A separate set of regressions is performed on the opnion scores considering non-trading days as well.

peaks in the time series of ambiguity index derived from the agent's minmax optimal actions also correspond distinctly to the recessionary episodes of the Lehman crisis in 2008 and the September 11 incident.

6 Conclusion

The paper makes a few key contributions. The first is it offers a *natural* Big Data and natural event method to measure an ambiguity adjustment through millions of actual textual posts in social and news media. The second key contribution is it proposes a binomial tree model of ambiguity based on the smooth ambiguity that models a representative agent's minimising behaviour to his perceived risks by controlling for information uncertainty. The model is able to explain salient features of investors' love for ambiguity during the recessionary times, and dislike for ambiguity during booming times, and also the higher value of the ambiguity premium during directional markets. Using a mixture of distributions model, the opinion scores from the financial and social media are also aggregated and an ambiguity index is formed. This ambiguity is distinct from disagreement although it is directly proportional to its square. It is also influenced by the individual information source uncertainty. The mixture of distributions model allows to test the relative importance of the social media and financial media news by varying its mixing weight parameter. In general, there is a higher weight on the importance of the financial news media ~ 0.55. However in recent times, especially during the Lehman and Greek debt crisis, the social media has affected asset prices greater. This importance could have resulted from the generally negativism of the social media during the recessionary periods and the maxmin utility behaviour of investor agent actions. The paper also offers a natural event evidence of agents' maxmin utility optimisation behaviour cited in Gilboa and Schmeidler (1989).

References

Ameriks, J. and Zeldes, S. P. (2000). How do household portfolio shares vary with age. *SSRN working paper*.

Bahra, B. (1997). Implied risk-neutral probability density functions from option prices: theory and application. *Bank of England Working Paper No 66*.

Baillon, A., Huang, Z., Selim, A., and Wakker, P. P. (2018). Measuring ambiguity attitudes for all (natural) events. *Econometrica*, 86.

Bekaert, G. and Wu, G. (2000). Asymmetric volatility and risk in equity markets. *Review of Financial Studies*, 13.

Bollen, J., Mao, H., and Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2:1–8.

Brenner, M. and Izhakian, Y. (2018). Asset pricing and ambiguity: Empirical evidence. *Journal of Financial Economics*, 130.

Brenner, M., Izhakian, Y., and Sade, O. (2015). Ambiguity and overconfidence. *SSRN*.

Campbell, J. Y. (1999). Asset prices, consumption and the business cycle. *Handbook of macroeconomics*, I.

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

Coles, J., Loewe, U., and Suay, J. (1995). On equilibrium pricing under parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 30.

Da, Z., Engelberg, J., and Gao, P. (2015). The sum of all fears investor sentiment and asset prices. *Review of Financial Studies*, 28:1–32.

Daniel, K. and Titman, S. (2006). Market reactions to tangible and intangible information. *Journal of Finance*, 61:1605–1643.

Epstein, L. G. and Schneider, M. (2008). Ambiguity, information quality, and asset pricing. *Journal of Finance*, 63.

Fruhwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

Ghysels, E., Anderson, E. W., and Juergens, J. L. (2009). The impact of risk and uncertainty on expected returns. *Journal of Financial Economics*, 94:233–263.

Gilboa, I. and Marinacci, M. (2013). *Ambiguity and the Bayesian Paradigm*. New York: Cambridge University Press, 2013.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18.

Heston, S. L. and Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts' Journal*, 73:67–83.

Illeditsch, P. K. (2011). Ambiguous information, portfolio inertia and excess volatility. *Journal of Finance*, 6:2213–2247.

Izhakian, Y. and Yermack, D. (2017). Risk, ambiguity, and the exercise of employee stock options. *Journal of Financial Economics*, pages 65–85.

Izhakian, Y., Yermack, D., and Zender, J. F. (2017). Ambiguity and the tradeoff theory of capital structure. *NBER working paper*.

Jiao, P., Veiga, A., and Walther, A. Social media, news media and the stock market. *SSRN*.

Kelly, B., Gentzkow, M., and Taddy, M. (2017). Text as data. *NBER working paper No 23276*.

Kilbanoff, P., Marinacci, M., and Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica*, pages 1849–1892.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries and 10-ks. *Journal of Finance*, 27.

Miao, J., Wei, B., and Zhou, H. (2019). Ambiguity aversion and the variance premium. *Quarterly Journal of Finance*, 09.

Nofsinger, J. and Wias, R. (1999). Herding and feedback trading by insitutional and individual investors. *Journal of Finance*, LIV.

Scruggs, J. T. (2002). Resolving the puzzling intertemporal relation between the market risk premium and conditional market variance: A two-factor approach. *Journal of Finance*, 53:575–603.

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

Tetlock, P. (2011). All the news that's fit to reprint: Do investors react to stale information? *Review of Financial Studies*, 24.

Veronesi, P. (2000). How does information quality affect stock returns? *Journal of Finance*, LV:807–837.