**Steps—Sentiment on Crash**

**「第一阶段」：互联网爬虫搜集资料的属性、平台**

语义解析引擎的建置需要大量的文本数据作为训练准确度之基底，故前期需要藉由爬虫的程序来进行网络理财投资的文本资料搜集的动作。本研究从五个不同的财经及投资平台网站搜集了上海50 ETF及深圳100 ETF股市的资料，并以2012年2月至2020年12月期间的日数据为样本期间。搜集的数据包括每日投资者在这两个股市所有相关财经网站平台对分别上市ETF个股发布的风险偏好、预期收益、观点及看涨及看跌等看法的充分信息，建置完整的股票投资人个股情绪的数据库。我们的最终样本涵盖了16010条来自这些ETF股票的公司新闻，爬取的资料来源于五个投资平台，分别是珞珈投资、中国网财经、新浪财经、同花顺财经和和讯股票。其中，珞珈投资是目前国内证券市场服务范围最为广泛的投资咨询公司，为一个各类投资者提供服务的交流平台。中国网则是国务院新闻办公室领导，中国外文出版发行事业局管理的国家重点新闻网站，是中国进行国际传播、信息交流的重要窗口。新浪财经、同花顺财经和和讯股票都是国内使用人数众多、知名度较高的股票交流平台。因此，本文综合考虑这五个财经平台发布的信息，对于衡量投资人情绪具有一定的代表性和公正性。

**「第二阶段」：结合网络社群聆听的语义挖掘和解析**

现如今应用语义挖掘所创造出来的价值，已超过过去以数据库搜寻为主的价值。在因特网以非常快速的方式，累积愈来愈多的非结构化数据的趋势下，实际运用语义挖掘来探索出各种未能预见的、创新的、重要的信息或知识，如商品评价、施政评价、民意调查、社群偏好等等，对政府机关、企业团体或个人而言，都是非常重要的工作。语义挖掘的相关研究和实际操作，已成为最近最为热门的话题。文本挖掘技术可用于过滤定性新闻报道中的噪声(Manning and Schutze, 2002)。与捕捉一般文本信息的文本分析相比，语义文本挖掘通常用于捕获语义特征文本，语义文本挖掘通常用于从大众信息中抽取有语义的特征词，考察支持某一事件或产品的不同面向的倾向。因此，从语义特征文本中获得的信息应该比从一般文本分析中获得的信息更精确。随着大量金融异常现象和非理性行为的出现，传统金融理论已经不能做出合理解释，学者们开始从新的角度解释金融异常现象，由此逐渐形成了行为金融理论。投资者情绪理论成为行为金融学的重要分支之一，而投资者情绪的测量问题一直是金融学研究的重点和难点之一，投资者情绪与股票收益之间的关系研究也一直是金融学研究的热点。本研究探讨了投资者情绪指数的构建以及其对股票市场收益的影响，对投资者的投资决策以及相关监管部门的监管具有重要的现实参考意义。

通过语义挖掘获得的数据，可以再透过语义解析引擎立即进行语义的判断并做出反馈，而语义解析引擎的建构须结合知识图谱的建置，建构投资人语义解析与财金讯息勾稽之知识图谱将会布署在云端等适合的环境中，最后再结合不同投资客群的社群聆听来解读投资大众对未来公司发展及公司股价涨跌直接与真实的评价，来与实际股价走势来相互印证。

**「第三阶段」：建立情绪词典**

我们使用一般类和金融类的情绪词典作为我们网络新闻的特征词的种子：一般类情绪词典的来源是台湾大学简体中文情感极性词典，而金融类的来源则是厦门大学中文财务情绪字典，它们都是使用中文字时投资者情绪的具有代表性和权威性的词典。前者的情感词典占我们特征词种子的40％，后者则占60％，因为我们的样本来自中国股票市场，并且投资者使用简体中文字。

**「第四阶段」：划分乐观及悲观情绪的术语**

我们使用特征词的种子来查找具有极度正向和负向分数的网络新闻作为我们潜在的特征词：我们首先将特征词的种子投入到我们150家*ETF*公司的所有网络新闻中。然后，我们采用流行的词频—逆文档频率（*TF-IDF*）方法，来计算和排序我们样本公司的每个新闻文档的正负情绪特征词的分数。

来自第*k*则新闻中乐观和悲观特征词的第*j*个术语的*TF-IDF*得分定义如下：



 (1)

例如=，表示包含乐观特征词的第*j*个术语的新闻文档的数量，且*N*是我们样本中总新闻文档的数目。然后，表示每个新闻文档中出现的乐观特征词的第*j*个术语的数量。和分别与和,有相似的定义。当新闻文件中的术语频率较高且所有新闻文档集合中的术语较少时，我们可以产生较高分数的*TF-IDF*。如果第*k*则新闻中的乐观特征词的得分高于悲观特征词的得分，则该新闻按乐观类别进行分类; 否则，该新闻按悲观类进行分类。我们选择具有前5％分数的新闻的特征词作为我们的正向特征词。同样，我们选择那些后5％得分的信息作为我们的负向特征词。

**「第五阶段」：检验乐观及悲观情绪术语的准确性**

一些特征词在乐观或悲观方面与其他特征词相比更具有相关性，我们将它们称为“代表性特征词”。 我们打性散所有网络新闻的正向和负向特征词的字数，并使用卡方检验来检验特定情感类的代表性特征词，并删除非情感的特征词（Frakes和Baeza-Yates，1992; Yang 和Pedersen，1997;和Lu等人，2013）。首先，虚无假设是第*j*个特征词不属于乐观类（包括悲观类和非情感类），而对立假设是第*j*个特征词属于乐观类。可以设计和检查类似但相反的虚无假设。 卡方独立测试的设计如下：

, (2)

其中，*A*和*B*分别表示第*L*类（例如，乐观类）中第*j*个和非第*j*个特征词的词频; *C*和*D*分别表示非*L*类（例如，非乐观类，包括悲观和非情感类）中的第*j*个和非第*j*个特征词。第*L*类中第*j*个特征词缺乏独立性意味着它可以鉴别为第*L*类的代表性特征词。 因此，第*j*个特征词可以被视为第*L*类中的代表性术语。 当有高卡方值时，独立性被拒绝。

**「第六阶段」：确定投资者情绪特征词的权重**

在选择具有代表性的特征词之后，应在语言领域确定其权重。我们延伸Oliveira等人（2016）计算*TF-IDF*修正方法的情绪权重,以显示在相关新闻文档中所挑选的第*j*项乐观词的权重。我们以挑选的乐观词中的第*j*项的权重为例进行以下计算：

 (3)

这里的下标*i*，*t*，*d*及*k*表示第*i*个公司在第*t*个月的第*d*天的第*k*则新闻；是第*k*则新闻文档中所挑选的乐观特征词中第*j*个词的权重。当较低时，乐观类中的第*j*个词较少；且当较高时，该词的词频在第*k*则新闻文档中也较高。因此，在第*k*则新闻文档中所挑选的乐观词中第*j*个词的权重较高，表明第*j*个乐观特征词在第*k*则新闻文档中更为重要。

表4列出了经过卡方独立检验后的代表性特征词以及相应的投资人乐观与悲观情绪的权重。悲观情绪的代表性特征词的数量（290）高于乐观情绪的代表性特征词的数量（165）。这种现象表明投资者对负面新闻具有过度反应（Zhu, Wu and Zhang and Yu, 2017）。我们基于等式（3）对新文档的第*j*个词求和并计算其相应的权重。正面或负面特征词的权重越高，表示这些特征词的乐观或悲观情绪的强度越高。

**「第七阶段」：The intensity of investor sentiment about individual stocks**

This step calculates the annual “intensity of investor sentiment” toward the *IOS* (*IOSit*) and *IPS* (*IPSit*) for *i*th firm at year *t*. To derive *IOSit* or *IPSit*, we mainly select the characteristic texts that investors are optimistic in the news for *i*th firm at year *t*. Taking investor optimistic sentiment for an example, we calculate *IOSit* by summing the daily intensity *IOSdk,it* as the sum of the term frequency  multiplied by weight in each firm at each year. Then, we calculate *IPSit* by summing daily influences of *IPSdk,it*. A similar approach is applied to calculate *IPSit*.

; (4)

, (5)

where  and  are the term frequencies of the *j*th characteristic text of investor optimistic and pessimistic sentiments for the *i*th firm on date *d* in the *k*th news;  and  are the weights of the *j*th characteristic texts of the optimistic and pessimistic sentiments. A higher intensity of *IOSit* or *IPSit* means a stronger optimistic or pessimistic sentiment of investors for firm *ith* on year *t*.

In addition, the maximum values of positive and negative sentiment variables are unlimited since the number of newsdocument *K* and number of characteristic texts *J* are unlimited. Thus, the concept of uniform distribution is used to standardize the nonzero-intensity values in which the minimum and maximum are between 1% and 100%, respectively.[[1]](#footnote-1) However, we still regard the zero-intensity value as zero. Thus, a high intensity means a strong investor sentiment.

Furthermore, to obtain out-of-sample classification accuracy, the real sentiment shown by investor message is compared with the sentiment score calculated using TF-IDF score. Meanwhile, the proportion of correct classification of dropping unclassified messages (CC), the proportion of correct classification of each class are computed. Also, to compare dictionary-base classification with machine learning classification, we use the machine learning techniques to estimate the sentiment score of individual messages that belongs to bull or bear class.

**「第八阶段」：Investor sentiment about stock market (X2)**

Some studies have used some indicators such as turnovers (TURN) (Jones and Owen, 2002; Baker and Wurgler, 2006) and discount rate of closed-end funds (CEFD) (Lee, Shleifer and Thaler, 1991) to measure the sentiments of investors on market (hereafter, market sentiment of investors). Other studies have used Ratio (i.e., the number of listed company’s going up divided by the number of listed company’s falling down) and Open (i.e., the growth rate of investors’opening accounts) to measure market sentiment of investors. IPO initial return is not a proper measure of investor sentiment although a few studies have considered as a sentiment proxy (Pitter and Welch, 2002; Baker and Wurgler, 2006). This is because there is always a high IPO initial return even if the market is not bullish. Thus, we employ the four market sentiment variables to create a comprehensive market sentiment index using the factor analysis method.



**「第九阶段」：**Adjusted Investor sentiment about pure individual stocks (X2)

The model that decomposes total loan herding into rational and irrational herding is as follows:



 (2)

**

** (3)

where  denotes investor sentiment about stock market and is defined by section 4.2.2; and  and  are the adjusted IOS and IPS measures in which we control investor sentiments about stock market, which denote the adjusted investor optimistic and pessimistic sentiments about pure individual stocks, respectively. Thus, the higher the  and  are, the stronger the investor optimistic and pessimistic sentiments about pure individual stocks are.

**「第十阶段」：投资者情绪对股票崩盤風險的影响和预测**

We first examine how investor optimistic and pessimistic sentiments about individual stocks affect the probability of stock price crash. Our model is specified as follows:

, (6)

where, the subscripts *i*, and *t* denote the *i*th stock at year *t*.

 (7)

1. For non-zero-intensity values, our standardized method is [99\*(*x* – *x*min)/(*x*max-*x*min) +1], where *x* is the intensity value. For zero-intensity values, we still use zero. [↑](#footnote-ref-1)