

年报文本语气能预示股价崩盘风险吗？

——基于 A 股上市公司年报的文本挖掘

摘 要

本文基于 2010 年至 2018 年 A 股上市公司年报的数据，研究年报文本语气对公司股价崩盘的影响。研究发现，年报文本净正向语气与公司下一年的股价崩盘风险呈显著的负相关关系。上述关系在年报存在不同方向的可操纵语气时表现不同。当年报文本存在负向的可操纵语气时，年报文本语气与股价崩盘风险显著负相关；而当年报文本存在正向可操纵语气时，年报文本语气对股价崩盘风险的影响不显著。研究表明，年报文本语气信息提供增量信息，且消极的语气信息更能引起市场的反应。本文揭示了年报文本语气与股价特征的联系，相关研究结论对深入年报非定量信息的研究，提高年报文本信息披露质量具有借鉴意义。

【关 键 词】年报 文本分析 股价崩盘

Does Annual Reports' Tone Foretell Stock Crash Risk?

--Based on Text Mining on Annual Reports of China A-share Companies

ABSTRACT

This paper studies the influence of annual reports' tone on companies' stock crash risk, based on data of China A-share companies from 2010 to 2018.

Results display that the annual report's net positive tone has the significantly negative relationship with company's next-year stock crash risk. And the relationship behaves differently when consider annual reports' tone manipulation. When the annual report contains the negative abnormal tone, the aforementioned relationship is significant; rather, when the annual report contains the positive abnormal tone, the aforementioned relationship is insignificant. This paper reveals that annual reports' tone provides extra

information, and the market reacts more notably on negative tone. The conclusion displays the relationship between annual reports' tone and stock crash risk, and has reference for further research on annual reports' non-quantitative information and for improving the quality of disclosure of textual information in annual report.

Key Words: Annual Report; Textual Analysis; Stock Crash Risk

本文档仅用于展示和学术探讨。

一、引言

年报是上市公司信息披露的重要途径，其中的信息对投资者而言意义非凡。年报的数据与文本都可能成为市场参与者进行投融资等决策的重要依据。年报信息可以大致分为两类，其一是以数据为主的定量信息，其二是以文本为主的定性信息。以往的研究主要集中于定量信息披露对市场的影响，且主要围绕着会计数据与经营数据展开。近年来随着计算机科学与大数据的突破性发展，利用大数据技术辅助分析年报文本成为了可能。因此，通过最新的技术对年报中的定性信息进行分析，能够从全新的维度补全对年报信息及其机制的研究，并加深对信息披露的理解。

自 2008 年金融危机之后，金融风险逐渐被各国重视，对股价崩盘的研究有其实际意义。学术界对股价崩盘的研究主要集中于公司治理层面，本文拟从年报信息披露角度对这一问题进行探究，试图揭示年报语气和股价崩盘之间的关联和传导机制。

本文通过研究 2010 年至 2018 年 A 股上市公司公开的年度报告，利用大数据文本分析与挖掘技术对年报中的非定量文本信息进行分析，以此来研究年报文本语气与公司股价崩盘风险的关系。本文认为，上市公司年报的文本语气是对股价崩盘风险的重要预测指标。年报文本净正向语气与股价崩盘风险呈现显著负相关关系，即年报的文本语气越消极，则公司在未来一年中股价崩盘的发生概率越大。此外，当年报存在负向的可操纵语气时，年报文本语气与股价崩盘风险的负相关关系更为显著，即年报语气主要通过负向的信息提升股价崩盘风险的方式进行。

本文可能的研究意义在于，丰富了已有的年报文本信息的作用以及股价崩盘影响因素的研究。本文还引入年报可操纵语气的概念，并对样本加以区分，以分析年报文本语气在不同方向可操纵语气的情境下对股价崩盘影响的不同。这丰富了相关领域的研究，有助于更好地理解上市公司信息披露中文本信息的重要性；同时本文也具有一定的现实意义，对于监管部门完善和监督上市公司信息披露，投资个体进行更全面地风险管理控制提供了重要的经验依据。

二、文献回顾与假设提出

（一）与年报文本信息有关的文献

早期的研究并不完全认可年报文本信息作为独立信息的影响与重要性。例如 Demers 和 Vega (2011) 利用文本挖掘算法研究指出，年报文本信息与年报中的会计利润表信息间有着复杂的关系：一些公司以为利润表没有表达足够的信息，会在文本中补充大量冗余信息，造成文本信息与会计利润表数据信息存在不同程度的内生性；另外，年报文本信息任由管理者意志决定，缺乏约束力，且其传达没有成本、信息难以核实，可以被视为“廉价交谈 (cheap talk)”。但随着研究渐趋深入，上市公司信息披露中文本信息的重要性逐渐被学术界所认识，越来越多的学者也接受年报文本信息有其特殊价值的观点。学界认为，年报文本信息是对上市公司数据化的盈余信息的重要补充与解释 (Muslu 等, 2015)，公司的业绩快报、年报与季报中的管理者讨论与分析 (Management Discussion and Analysis, 即 MD&A) 部分等文本信息能引起市场显著的短期反应 (Davis 等, 2012)。

在研究对象的选择上，现有研究主要围绕在年报文本信息（尤其是 MD&A 部分）和业绩说明会的文本信息。年报文本信息能引起市场的短期反应，尤其是分析师的预期。分析师会关注年报文本的信息并据此调整对上市公司的业绩预测，同时年报文本信息也能引起股价的波动 (许文瀚和朱朝晖, 2019)。年报中的语气与上市公司未来的业绩有关系。非财务信息会影响外部融资，进而对企业整体投资效率产生影响 (宋旭霞, 2016)。薛爽等 (2010) 的研

究在控制了相关财务指标之后，发现年报 MD&A 提及的原因及其特征与下一年公司能否扭亏为盈的概率有显著关联。

对业绩说明会文本信息中的管理层语调研究也得出了类似的结论。业绩说明会源于 2004 年深交所对中小企业发布的《深圳证券交易所中小企业板块上市公司诚信建设指引》，规定中小板、创业板上市公司应每年举行年度报告说明会。现在也有许多主板上市公司为树立自身形象也会参与其中。投资者会对业绩说明会中管理层的语调作出显著的同向反应（林乐和谢德仁，2016），且管理层的语调与公司在下一年的业绩显著相关（谢德仁和林乐，2015）。管理层在业绩说明会中的积极语义信息会通过提振投资者情绪的方式，化解未来股价崩盘的风险（杨七中等，2019）。

正因为文本信息能够产生如此影响，且语言犯错不存在规律的模式进行判断与参照，语言或文本披露要比定量披露存在更大的操作空间（Davis 等，2012）。这种的语言灵活性使得管理者操纵文本内容以达到自身目的。现有研究已揭示，管理者会操纵文本信息，且对年报文本语气的操纵往往与应计操纵同时进行并辅助管理层进行盈余管理（王华杰和王克敏，2018）。管理者还会操纵文本信息的复杂性为自己获取更高的超额薪酬并提高公司估值（王克敏等，2018）。且当公司盈余管理空间较小时，管理者更可能操纵文本语气（Huang 等，2014）。

此外，企业信息披露中的负面语调是重要的风险指标。MD&A 中的前瞻性信息能够显著提升外界对上市公司财务危机的预测能力（李秉成等，2019）。MD&A 文本中“净负面语调”比例越高，说明企业的信用风险越大，企业也会面临越严格的合同条款（Ertugrul 等，2017）。

（二）与股价崩盘有关的文献

股价崩盘是股价突然大幅下降的经济现象。主流观点认为，股价崩盘的形成机理如下：管理层成功地隐藏坏消息存在一个临界点，当经理隐藏坏消息的成本高到这个临界点时，坏消息会突然全部释放出来，从而股票价格急剧下滑，导致崩盘（Jin 和 Myers，2006）。潘越等（2011）的研究表明上市公司透明度越低，股价崩盘风险越高，而分析师关注可以减弱信息不透明对股价崩盘风险的影响。许年行等（2012）发现，分析师乐观偏差增加了未来股价崩盘风险。也有研究从企业内部角度对股价崩盘风险进行探究。目前的研究表明，税务机关较高的税收征管的力度、媒体的关注程度、审计师和分析师良好的声誉及报告、乐观的投资者态度等外部市场因素会抑制股价崩盘风险的产生，在公司内部的层面，当公司拥有较为完善的内部控制与良好的企业社会责任意识（叶康涛等，2015）、较高的大股东持股比例（王化成等，2015）、良好的会计稳健性（Kim 等，2019）和会计报告信息透明度（Hutton 等，2009）等时，公司能够降低股价崩盘风险。

（三）年报文本语气与股价崩盘风险关系

年报文本语气能否影响公司股价的表现，进而预示股价崩盘风险？本文认为，年报作为重要的信息披露文件，无论是其定量信息还是定性信息均承载着一定的价值，都能或多或少地影响投资者的判断与决策。年报的文本语气能够为投资者带来额外的信息增量（Feldman 等，2010），进而通过投资者的行为投射到股价上。

年报的文本语气能够影响分析师和市场投资者的预测。年报的文本能够帮助投资者进行决策，例如，MD&A 文本信息内容和语言语调特征有助于使用者提高盈余预测的准确度

（Bochkay 和 Levine，2019）。但是，年报文本语气更有可能被操纵，这使得年报文本有背离真实情况的可能。已有研究证实了管理者会操纵文本信息，操纵文本信息的复杂性（王克敏

等, 2018), 和文本的语气信息, 以此提高信息的解读成本, 并由此引导市场作出对管理层的能力和 Company 价值有利的判断。此外, 年报文本语气往往与应计操纵结合着进行, 而且当公司在应计操纵约束较大、跟踪分析师较多、机构持股比例更高时, 管理者更可能操纵文本语气 (王华杰和王克敏, 2018)。股价崩盘风险往往与对信息的操纵有关, 因为股价崩盘是由积累到一定程度之后释放而出的负面信息引发的 (Jin 和 Myers, 2006)。基于此, 本文提出以下待检验研究假说 H1:

H1: 在其他条件相同的情况下, 年报文本的语气与股价崩盘风险呈负相关关系。

Huang 等 (2014) 的研究关注了年报文本语气操纵, 在其研究中将年报文本语气中超出可预期的部分界定为“可操纵语气”(abnormal tone), 并将其视为管理者操纵年报文本语气的手段之一。研究亦表明, 可操纵语气能够影响证券的估值。由于投资者会参照年报及估值等信息进行投资决策, 且行为金融学研究已经揭示了投资者的损失厌恶情绪, 相较于积极的语气信息, 人们对消极的信息反应更为显著, 因此管理者对年报语气的操纵行为可能对年报语气与股价崩盘的预示作用有调节效应。市场对不同方向的年报文本语气操纵反应也应当有所不同。

正向的操纵含义与影响不甚明朗, 因为如果存在正向的语气操纵, 既有可能是管理层对已有功劳的夸大, 也有可能是为了掩盖不良信息而刻意为之。也就是说, 一方面中小投资者对文本信息过度依赖 (Baginski 等, 2016), 这使得管理者通过操纵年报的文本信息, 以期提高公司市场估值 (王克敏等, 2018)。上市公司自愿披露发展信息有助于降低分析师预测误差, 缓解信息不对称 (Muslu 等, 2011)。但是另一方面, 积极的语调可能是管理层进行印象管理的结果, 并非是对公司前景的看好 (周波等, 2019), 这一现象在真实程度越低的情况下越显著。而在短期, 市场会对年报中的语调产生反应: 投资者会对业绩说明会、年报 MD&A 部分等的正向语调做出显著的正向反应 (林乐和谢德仁, 2016), 分析师也会对管理层的净正面语调形成显著的正向反应 (林乐和谢德仁, 2017)。此时, 正向的操纵会加大市场上的乐观偏差, 可能在未来积累更多的股价崩盘风险。

但另一面, 负向的操纵可能会加深投资者的避险情绪。管理层语调具有信息增量, 且投资者与分析师都有“听话听音”的习惯, 会从语气中获取整体的感知并改变自身投资决策 (林乐和谢德仁, 2016)。管理层被认为是公司的内部人士, 掌握比公司外部人员更多的内部消息, 因此市场往往对其负面言论更为关心。管理层净正面语调与公司下一年业绩显著正相关 (林乐和谢德仁, 2015), 如果管理层有负向的语气操纵, 可能会被外部人士解读为管理层自身对未来都不抱希望, 没有信心, 这会加剧投资者与市场对公司未来业绩的失望情绪与态度。此外, 管理者在年报中的语调操纵能与盈余管理形成配合作用, 都能起到掩盖管理者动机的目的, (朱朝晖和许文瀚, 2018), 会计稳健性与透明度由此被降低。因此当年报中存在负向的语气操纵时, 可能引起市场对管理者自身的消极态度, 以及会计信息质量的质疑, 这可能会积累股价崩盘风险, 甚至直接触发临界点导致崩盘发生。

综上, 本文提出以下待检验研究假说 H2:

H2: 在其他条件相同的情况下, 当年报可操纵语气为负时, 年报文本语气与股价崩盘的负相关关系比年报可操纵语气为正时更显著。

三、样本选择与数据来源

(一) 数据来源

本文选取了 2010 年至 2018 年 A 股所有上市公司的年报及市场价格信息为研究样本, 并对样本进行了如下的筛选: (1) 删除了因爬虫软件提取有误而导致文本长度过短 (中文字数少于 20000 个字) 的年报样本; (2) 删除所有者权益小于等于零的样本; (3) 删除年报公布

后 1 年内交易周数少于 30 的样本；(4) 删除相关数据缺失的样本；(5) 删除相关数据无法计算的样本；(6) 对所有的样本进行上下 1% 的缩尾。本文最终获得了 18203 个包括了 3350 家上市公司年报的样本作为对假设一进行实证检验的样本集，获得了 11687 个包括了 2375 家上市公司年报的样本作为对假设二的样本集。上述两个样本集样本总数的差异是在对下文公式 (5) 进行计算时相关数据缺失导致的。

本文的年报数据来源于中财网 (<http://www.cfi.cn>)，本文运用 python 编程进行爬虫获取该网站上所有可获取的年报文本数据，进行数据清洗格式调整，并只保留其中的中文字符，而后通过 python jieba 库对年报文本信息进行分析。本文的公司财务数据、股票交易数据和年报审计数据等均来自 CSMAR 国泰安数据库。

(二) 变量定义

1. 因变量：股价崩盘指标

本文参考许年行等 (2012)、杨七中等 (2019) 的研究，通过构建 $NCSKEW_{i,t}$ 和 $DUVOL_{i,t}$ 两个指标衡量上市公司的股价崩盘风险。为了更准确地衡量年报发布之后年报语气对股价崩盘的影响，本文的股价崩盘指标计算的是自年报发布日之后 1 年之内交易周的股价崩盘风险。

首先，本文对每家上市发布年报后 1 年内公司股票 i 的周收益数据进行回归计算：

$$r_{i,t} = \alpha + \beta_1 r_{m,t+2} + \beta_2 r_{m,t+1} + \beta_3 r_{m,t} + \beta_4 r_{m,t-1} + \beta_5 r_{m,t-2} + \varepsilon_{i,t} \quad (1)$$

其中 $r_{i,t}$ 表示股票 i 第 t 周的收益率， $r_{m,t}$ 表示市场第 t 周的收益率，本文在模型中加入了收益率的超前 2 项和滞后 2 项，以期修正并尽可能地减少因为股票的非同步交易而导致的计量误差。 $\varepsilon_{i,t}$ 表示回归模型的残差，即股票收益率中不能被市场收益率波动所解释的部分。本文设 $W_{i,t} = \ln(1 + \varepsilon_{i,t})$ ，表示股票 i 在第 t 周的公司特有收益率。

其次，按照公式 (2) 和 (3)，使用 $W_{i,t}$ 计算股价崩盘风险度量指标负收益偏态系数和股票收益上下波动率。

首先是负收益偏态系数 ($NCSKEW_{i,t}$)：

$$NCSKEW_{i,t} = -[n(n-1)^{3/2} \sum W_{i,t}^3] / [(n-1)(n-2)(\sum W_{i,t}^2)^{3/2}] \quad (2)$$

其中， n 为股票 i 的上市公司在发布年报后的 1 年内上市交易的周数， $NCSKEW_{i,t}$ 即负收益偏态系数，其数值越大，说明该公司股价在第 t 期年报发布后的 1 年时间内，股价向负向偏斜的频率越高，即股价崩盘的可能性也越大。

其次是股票收益上下波动率 ($DUVOL_{i,t}$)：

$$DUVOL_{i,t} = \log\{[(n_{up} - 1) \sum_{Down} W_{i,t}^2] / [(n_{down} - 1) \sum_{Up} W_{i,t}^2]\} \quad (3)$$

其中， n_{up} 表示股票 i 在发布年报后 1 年内的周特有收益率 $W_{i,t}$ 高于该股票年平均收益率均值 $\overline{W_{i,t}}$ 的周数， n_{down} 表示股票 i 在发布年报后 1 年内的周特有收益率 $W_{i,t}$ 低于该股票年平均收益率均值 $\overline{W_{i,t}}$ 的周数。 $DUVOL_{i,t}$ 即股票收益上下波动率，其数值越大，说明股票 i 在第 t 期年报发布后的 1 年内收益向下波动的幅度与向上波动的幅度之比越高，即向下波动更容易发生，更有可能出现股价崩盘的情况。

2. 自变量：文本语气信息

本文借鉴了相关文献 (Huang 等, 2014; 王克敏等, 2018) 的研究方法进行年报文本语气 ($TONE_t$) 的计算。本文的语气词库使用了台湾大学《中文情感极性词典》、台湾国立大学《简体中文情感词典》和知网 Hownet 情感词典。这些词典将词汇划分为积极属性词集和消极属性词集，本文取这些词典词汇的并集，即将两种属性的词汇分别按词集进行拼合后删去重复的部分，最终获得了 5127 个积极词汇和 8941 个消极词汇作为本文语气判定的词集。另外，为防止分词算法将诸如“投资损失”“负债合计”等会计术语误判为情感词汇，本文整理了

常用且与本文整理词集相重合的会计术语 306 个，作为 jieba 分词的额外停用词表。当分词过程中出现额外停用词表中的术语时，算法会自动跳过该词，而不会将其统计成为积极词汇或消极词汇。限于篇幅，积极词汇词集、消极词汇词集和会计术语额外停用词表不在文中展示，相关表格备索。

本文首先调用 jieba 库对每份年报文本数据进行分词，而后对分词结果中的每个词汇进行词频统计并判定是否属于积极、消极词库，以此计算年报中的积极词数量（POS）与消极词数量（NEG），再按照公式(4)计算年报文本语气：

$$TONE_{i,t} = (POS - NEG) / (POS + NEG) \quad (4)$$

同时，本文还借鉴了 Huang 等（2014）和王华杰等（2018）的计量方法计算了年报文本可操纵语气（ $ABTONE_{i,t}$ 和 $ABTONE_FE_{i,t}$ ），以此衡量管理层对年报文本语气的操纵程度。本文使用如下模型分别计算未控制和控制未来盈利状况的年报文本可操纵语气：

$$TONE_{i,t} = \alpha + \beta_1 ROA_{i,t} + \beta_2 STD_ROA_{i,t} + \beta_3 RET_{i,t} + \beta_4 STD_RET_{i,t} + \beta_5 SIZE_{i,t} + \beta_6 BM_{i,t} + \beta_7 AGE_{i,t} + \beta_8 LOSS_{i,t} + \beta_9 D_EARN_{i,t} + \varepsilon_{i,t} \quad (5)$$

$$TONE_{i,t} = \alpha + \beta_1 ROA_{i,t} + \beta_2 STD_ROA_{i,t} + \beta_3 RET_{i,t} + \beta_4 STD_RET_{i,t} + \beta_5 SIZE_{i,t} + \beta_6 BM_{i,t} + \beta_7 AGE_{i,t} + \beta_8 LOSS_{i,t} + \beta_9 D_EARN_{i,t} + \beta_{10} ROA_{i,t+1} + \varepsilon_{i,t} \quad (6)$$

模型中， $ROA_{i,t}$ 即 t 期总资产回报率； $STD_ROA_{i,t}$ 即公司过去五年总资产回报率的标准差； $RET_{i,t}$ 是公司 t 期所有交易周收益率的平均值； $STD_RET_{i,t}$ 是公司 t 期股票所有交易周收益率的标准差； $SIZE_{i,t}$ 是 t 期总资产的自然对数； $BM_{i,t}$ 是 t 期公司账面市值比； $AGE_{i,t}$ 是公司 t 期上市年数的自然对数； $LOSS_{i,t}$ 是判断公司 t 期是否亏损的哑变量，若当年亏损则为 1，其他为 0； $D_EARN_{i,t}$ 是公司 t 期净利润减 $t-1$ 期净利润，除以 $t-1$ 期总资产的比值。按年度对公式(5)进行回归，所得到的残差为公司年报文本可操纵语气 $ABTONE_{i,t}$ 。公式(6)在公式(5)的基础上增加了 $ROA_{i,t+1}$ 即 $t+1$ 期总资产回报率，回归残差为控制了固定效应的年报文本可操纵语气 $ABTONE_FE_{i,t}$ 。本文使用公式（5）计算而出的 $ABTONE_{i,t}$ 进行对假设二的实证检验，并使用公式（6）的 $ABTONE_FE_{i,t}$ 再进行检验，以期提高本文研究的稳健性。

3. 控制变量

本文综合考虑已有对股价崩盘风险和年报文本语气相关文献（许年行等，2012；An 和 Zhang，2013；叶康涛等，2015）和所研究问题的具体情况，选取如下控制变量：周收益率平均值、周收益率标准差、超额换手率、企业规模、资产负债率、总资产收益率、账市比、可操纵应计程度和第一大股东的持股比例等。另外，本文还控制了年度效应和行业效应。相关变量定义及名称详见表 1。

表1：变量名称及定义

变量名称	变量符号	变量定义
负收益偏态系数	$NCSKEW_{i,t}$	详见变量定义
股票收益上下波动率	$DUVOL_{i,t}$	详见变量定义
年报文本净正向语气	$TONE_{i,t}$	详见变量定义
年报文本可操纵语气	$ABTONE_{i,t}$	详见变量定义
年报文本可操纵语气， 控制未来盈利状况	$ABTONE_FE_{i,t}$	详见变量定义
年度周收益率平均值	$RET_{i,t}$	公司 t 期所有交易周收益率的平均值
年度周收益率标准差	$SIGMA_{i,t}$	公司 t 期股票所有交易周收益率的标准差
年度周换手率平均值	$TURNOVER_{i,t}$	公司 t 期股票所有交易周换手率的平均值
总资产收益率	$ROA_{i,t}$	公司 t 期总资产回报率
资产负债率	$LEV_{i,t}$	公司 t 期资产负债率
账面市值比	$BM_{i,t}$	公司 t 期账面市值比
企业规模	$SIZE_{i,t}$	公司 t 期总资产的自然对数
大股东持股比例	$HOLD_{i,t}$	公司 t 期第一大股东的持股比例
可操纵应计	$ACCM_{i,t}$	按照修正琼斯模型计算的可操纵应计，分组同行业哑变量

年报审计意见	$OPINION_{i,t}$	公司 t 期年报审计意见哑变量，若注册会计师出具标准无保留意见取值为1，否则为0
年报审计质量	$AUDITOR_{i,t}$	公司 t 期年报审计质量哑变量，若审计的事务所为国际四大会计事务所以取值为2，若为国内八大国资事务所以取值为1，其余为0
行业哑变量	$industry_{i,t}$	采用证监会2012年分类标准，制造业取前两位字符，其余取第一位字母
年度哑变量	$year_{i,t}$	2010~2017年

四、实证分析

（一）描述性统计与相关性检验

本文对所有的连续变量进行了缩尾处理（winsorize），即在连续变量的上下1%水平上删去极端值。极值处理后的变量描述性统计如下表2所示。其中为了方便阅读，将 $ABTONE_{i,t}$ 与 $ABTONE_FE_{i,t}$ 的值在原始值的基础上乘以100加以展示。即 $ABTONE_{i,t}$ 的原始值均值和中位数分别为0.000371和0.001042， $ABTONE_FE_{i,t}$ 的原始值均值和中位数分别为0.00038和0.001077。

通过表2可以得知， $TONE_{i,t}$ 的均值为0.4832，中位数为0.4962，标准差为0.0904；而 $ABTONE_{i,t}$ 和 $ABTONE_FE_{i,t}$ 的标准差分别为6.6262和6.6209。 $ABTONE_{i,t}$ 和 $ABTONE_FE_{i,t}$ 的均值分别为0.0371和0.0380，中位数分别为0.1042和0.1077。 $ABTONE_{i,t}$ 的1%和99%分位数的绝对值分别为15.9844和15.9071，其25%和75%分位数的绝对值分别为4.1152和4.3759； $ABTONE_FE_{i,t}$ 的1%分位数和99%分位数的绝对值分别为15.9935和15.8828，其25%和75%分位数的绝对值分别为4.1213和4.3690。分位数的数据说明从整体而言，样本公司的年报文本既有向上的可操纵语气也有向下可操纵语气，整体而言略向上的语气操纵较多，向上与向下的语气操纵空间幅度基本对称，而且分布大体一致。

表2：样本主要变量描述性统计

变量名	Obs.	平均值	中位数	标准差	分位数			
					1%	25%	75%	99%
$TONE_{i,t}$	18203	0.4832	0.4962	0.0904	0.2468	0.4368	0.5438	0.6565
$ABTONE_{i,t}$	11687	0.0371	0.1042	6.6262	-15.9844	-4.1152	4.3759	15.9071
$ABTONE_FE_{i,t}$	11687	0.0380	0.1077	6.6209	-15.9935	-4.1213	4.3690	15.8828
$DUVOL_{i,t+1}$	18203	-0.2391	-0.2403	0.4746	-1.3112	-0.5655	0.0839	0.8877
$NCSKEW_{i,t+1}$	18203	-0.3416	-0.3085	0.7186	-2.2414	-0.7647	0.0978	1.4380
$RET_{i,t}$	18203	0.0018	0.0003	0.0100	-0.0155	-0.0051	0.0070	0.0339
$SIGMA_{i,t}$	18203	0.0635	0.0579	0.0252	0.0282	0.0469	0.0726	0.1512
$TURNOVER_{i,t}$	18203	-0.0162	-0.0082	0.0841	-0.2845	-0.0528	0.0258	0.1764
$ROA_{i,t}$	18203	0.0429	0.0383	0.0525	-0.1265	0.0151	0.0698	0.1844
$LEV_{i,t}$	18203	0.4371	0.4315	0.2066	0.0660	0.2700	0.5954	0.8673
$BM_{i,t}$	18203	0.6084	0.6097	0.2329	0.1427	0.4268	0.7910	1.0643
$SIZE_{i,t}$	18203	22.16	21.98	1.26	19.95	21.25	22.86	25.88
$HOLD_{i,t}$	18203	0.3488	0.3300	0.1499	0.0877	0.2314	0.4500	0.7409
$ACCM_{i,t}$	18203	0.0032	0.0001	0.1000	-0.2833	-0.0455	0.0484	0.3154
$OPINION_{i,t}$	18203	0.9754	1.0000	0.1549	0.0000	1.0000	1.0000	1.0000
$AUDITOR_{i,t}$	18203	0.7172	1.0000	0.5553	0.0000	0.0000	1.0000	2.0000

对所有的连续变量进行皮尔森相关性检验，检验结果如附表1所示。由于表格较大，检验结果（附表1）被放置在本文文末。根据检验数据， $TONE_{i,t}$ 与 $NCSKEW_{i,t+1}$ 和 $DUVOL_{i,t+1}$ 的相关性分别为-0.046和-0.072，均在1%水平上显著。这说明年报文本的语气与公司的在下一

一年度的股价崩盘风险存在显著的负相关关系。当年报中披露了更多的积极语气时，公司股价崩盘的风险就越小，这初步印证了本文的假设一。

此外，本文对所有变量的方差膨胀系数（VIF）进行了检验，发现均小于临界值 5。这说明本文基本不存在多重共线性的情况。方差膨胀系数结果如表 3 所示。

表3：样本主要变量方差膨胀系数

变量名	VIF	变量名	VIF
$TONE_{i,t}$	1.0443	$TONE_{i,t}$	1.0483
$NCSKEW_{i,t+1}$	1.0285	$DUVOL_{i,t+1}$	1.0380
$RET_{i,t}$	1.7947	$RET_{i,t}$	1.7988
$SIGMA_{i,t}$	2.0316	$SIGMA_{i,t}$	2.0320
$TURNOVER_{i,t}$	1.2711	$TURNOVER_{i,t}$	1.2713
$ROA_{i,t}$	1.4438	$ROA_{i,t}$	1.4438
$LEV_{i,t}$	1.7187	$LEV_{i,t}$	1.7191
$BM_{i,t}$	1.9839	$BM_{i,t}$	1.9796
$SIZE_{i,t}$	2.0731	$SIZE_{i,t}$	2.0736
$HOLD_{i,t}$	1.0599	$HOLD_{i,t}$	1.0601
$ACCM_{i,t}$	1.0510	$ACCM_{i,t}$	1.0510
$OPINION_{i,t}$	1.0517	$OPINION_{i,t}$	1.0515
$AUDITOR_{i,t}$	1.0757	$AUDITOR_{i,t}$	1.0756
常数项	594.5677	常数项	595.3231

（二）多元回归检验结果与分析

为了控制其他因素对年报文本语气的影响，本文设计了公式（7）和（8）以下的模型进行多元回归分析，检验年报文本语气对股价崩盘情况的影响。本文在模型中控制了年度与行业的固定效应，因此本文以下实证结果的标准误均在行业层面进行了聚类处理。

$$NCSKEW_{i,t+1} = \alpha + \beta_1 TONE_{i,t} + \beta_2 RET_{i,t} + \beta_3 SIGMA_{i,t} + \beta_4 TURNOVER_{i,t} + \beta_5 ROA_{i,t} + \beta_6 LEV_{i,t} + \beta_7 BM_{i,t} + \beta_8 SIZE_{i,t} + \beta_9 HOLD_{i,t} + \beta_{10} ACCM_{i,t} + \beta_{11} OPINION_{i,t} + \beta_{12} AUDITOR_{i,t} + \varepsilon_{i,t} \quad (7)$$

$$DUVOL_{i,t+1} = \alpha + \beta_1 TONE_{i,t} + \beta_2 RET_{i,t} + \beta_3 SIGMA_{i,t} + \beta_4 TURNOVER_{i,t} + \beta_5 ROA_{i,t} + \beta_6 LEV_{i,t} + \beta_7 BM_{i,t} + \beta_8 SIZE_{i,t} + \beta_9 HOLD_{i,t} + \beta_{10} ACCM_{i,t} + \beta_{11} OPINION_{i,t} + \beta_{12} AUDITOR_{i,t} + \varepsilon_{i,t} \quad (8)$$

表 4 是使用公式（7）（8），对于假设一的多元回归分析结果。表 4 的列（1）显示，使用 $NCSKEW_{i,t+1}$ 作为因变量时， $TONE_{i,t}$ 的系数为-0.413，在 1% 的水平上显著；列（2）显示，使用 $DUVOL_{i,t+1}$ 作为因变量时， $TONE_{i,t}$ 的系数为-0.428，也在 1% 的水平上显著。结果说明，年报文本语气是有价值的非定量信息，仍然能为投资者提供增量信息，减小投资者与公司之间的信息不对称情况，进而影响公司的股价表现。年报文本语气与公司在下一年度的股价崩盘风险确实存在负相关的关系，即年报披露了更多的消极语气，年报文本净正向语气越低，下一年度的股价崩盘风险越高。

在控制变量方面， $RET_{i,t}$ 与股价崩盘风险指标显著负相关（罗进辉和杜兴强，2014）。 $HOLD_{i,t}$ 与股价崩盘指标显著负相关，表明大股东持股显著存在监督效应（王化成等，2015）。这与以往的研究结果相一致。

表4：年报文本净正向语气与股价崩盘指标回归结果

	(1)	(2)
	$NCSKEW_{i,t+1}$	$DUVOL_{i,t+1}$
$TONE_{i,t}$	-0.413*** (0.098)	-0.428*** (0.084)

$RET_{i,t}$	-3.168*	-4.078***
	(1.623)	(0.972)
$SIGMA_{i,t}$	-6.773***	-4.065***
	(0.561)	(0.294)
$TURNOVER_{i,t}$	-0.533***	-0.367***
	(0.099)	(0.064)
$ROA_{i,t}$	0.059	0.064
	(0.124)	(0.082)
$LEV_{i,t}$	0.088**	0.072***
	(0.041)	(0.025)
$BM_{i,t}$	-0.334***	-0.173***
	(0.052)	(0.037)
$SIZE_{i,t}$	-0.006	-0.013
	(0.012)	(0.008)
$HOLD_{i,t}$	-0.086**	-0.076***
	(0.035)	(0.024)
$ACCM_{i,t}$	0.069	0.035
	(0.066)	(0.050)
$OPINION_{i,t}$	-0.133***	-0.075***
	(0.039)	(0.027)
$AUDITOR_{i,t}$	-0.004	0.004
	(0.012)	(0.008)
(年份)	控制	控制
(行业)	控制	控制
常数项	0.824***	0.720***
	(0.271)	(0.185)
Obs.	18203	18203
R-squared	0.056	0.060

*** p<0.01, ** p<0.05, * p<0.1

为了检验假设二，本文按照年报文本可操纵语气的方向对样本进行分组。具体地，将每个样本的 $ABTONE_{i,t}$ 取值与 0 进行比较，当 $ABTONE_{i,t}$ 的取值大于等于 0 时，将该样本归于“正向组”；当 $ABTONE_{i,t}$ 的取值小于 0 时归入“负向组”。如果该样本的 $ABTONE_{i,t}$ 因为数据缺失无法计算，则删除该样本。经过分组后，获得了 11687 个样本，其中“正向组”样本 6019 个，“负向组”样本 5668 个。

使用公式（7）（8）分别对“正向组”与“负向组”进行多元回归。表 5 是对于假设二的分组多元回归结果。其中列（1）（3）使用 $NCSKEW_{i,t+1}$ 作为因变量，列（2）（4）使用 $DUVOL_{i,t+1}$ 作为因变量。列（1）（2）的结果显示，当年报文本可操纵语气为非负时，年报文本语气的回归系数分别为-0.066 和-0.133，t 值分别为 0.253 和 0.155，均不显著。即当年报文本语气存在向积极方向操纵的空间时，年报文本语气对下一年股价崩盘的影响不显著。列

（3）（4）的结果显示，当年报文本可操纵语气为负时，年报文本语气的回归系数分别为-0.378 和-0.528，系数的绝对值大于列（1）（2）正向组的系数绝对值，且列（3）（4）的结果分别在 10%的水平与 1%的水平上显著。即当年报文本语气存在向消极方向操纵的空间时，年报文本语气对下一年股价崩盘的影响显著为负。分别对比列（1）和（3），列（2）和（4）的回归结果可以发现，在控制了其他因素的情况之下，存在消极方向年报文本可操纵语气的年报对股价崩盘的影响更为显著。可见，当投资者发现年报文本语气存在消极方向的操纵空间时，即意识到管理层存在负向文本语气操纵时，出于避险情绪投资者会对年报文本语气更为关注，投资者一方面认为该上市公司的信息透明度不足，另一方面对年报中的文本语气信息持更为悲观的态度，公司在下一年的股价崩盘风险由此提升。

表5：年报文本净正向语气与股价崩盘指标回归结果——以年报可操纵语气方向分组

	$ABTONE_{i,t} \geq 0$		$ABTONE_{i,t} < 0$	
	(1)	(2)	(3)	(4)
	$NCSKEW_{i,t+1}$	$DUVOL_{i,t+1}$	$NCSKEW_{i,t+}$	$DUVOL_{i,t+1}$
$TONE_{i,t}$	-0.066 (0.253)	-0.133 (0.155)	-0.378* (0.227)	-0.528*** (0.160)
$RET_{i,t}$	-1.541 (2.219)	-3.083** (1.289)	-7.443*** (2.408)	-6.093*** (1.466)
$SIGMA_{i,t}$	-6.811*** (0.874)	-3.667*** (0.490)	-5.553*** (0.933)	-3.398*** (0.498)
$TURNOVER_{i,t}$	-0.561** (0.234)	-0.380*** (0.136)	-1.005*** (0.315)	-0.692*** (0.191)
$ROA_{i,t}$	-0.150 (0.239)	-0.161 (0.144)	0.179 (0.229)	0.220 (0.168)
$LEV_{i,t}$	0.146** (0.057)	0.072** (0.036)	0.098 (0.066)	0.065 (0.039)
$BM_{i,t}$	-0.420*** (0.069)	-0.248*** (0.045)	-0.365*** (0.084)	-0.160*** (0.058)
$SIZE_{i,t}$	-0.004 (0.013)	-0.004 (0.009)	0.008 (0.018)	-0.006 (0.011)
$HOLD_{i,t}$	-0.135** (0.064)	-0.093** (0.044)	-0.134* (0.074)	-0.142*** (0.047)
$ACCM_{i,t}$	0.075 (0.123)	0.007 (0.087)	0.154 (0.109)	0.043 (0.081)
$OPINION_{i,t}$	-0.209*** (0.079)	-0.120** (0.047)	-0.068 (0.062)	-0.042 (0.046)
$AUDITOR_{i,t}$	-0.007 (0.022)	0.000 (0.014)	-0.008 (0.018)	-0.002 (0.011)
(年份)	控制	控制	控制	控制
(行业)	控制	控制	控制	控制
常数项	0.756** (0.325)	0.490** (0.214)	0.387 (0.421)	0.560** (0.267)
Obs.	6019	6019	5668	5668
R-squared	0.050	0.046	0.066	0.071

*** p<0.01, ** p<0.05, * p<0.1

(三) 稳健性检验与进一步研究

1. 替换分组依据

为了检验假设二的稳健性，本文采用控制未来盈利能力的 $ABTONE_FE_{i,t}$ 作为分组依据，重复上述分组回归的过程进行实证检验。按照 $ABTONE_FE_{i,t}$ 的正负性将样本划分为“正向组”与“负向组”并分别进行多元回归，回归结果如表6所示。对比列（1）和（3），可知 $TONE_{i,t}$ 的回归系数分别为-0.200和-0.605，其中正向组不显著而负向组在1%的水平上显著。对比（2）和（4）可知 $TONE_{i,t}$ 回归系数分别为-0.291和-0.697，正向组仍不显著，负向组在1%的水平上显著。由此得出了相同的结论，即当年报文本存在负向操纵空间时，年报文本语气越消极，股价崩盘风险越大。

表6：年报文本净正向语气与股价崩盘指标回归结果——以替代

	指标分组			
	$ABTONE_FE_{i,t} \geq 0$		$ABTONE_FE_{i,t} < 0$	
	(1)	(2)	(3)	(4)
	$NCSKEW_{i,t+1}$	$DUVOL_{i,t+1}$	$NCSKEW_{i,t+1}$	$DUVOL_{i,t+1}$
$TONE_{i,t}$	-0.200 (0.282)	-0.291 (0.178)	-0.605*** (0.212)	-0.697*** (0.151)
(控制变量)	控制	控制	控制	控制
(年份)	控制	控制	控制	控制
(行业)	控制	控制	控制	控制
常数项	1.100*** (0.309)	0.743*** (0.217)	1.008*** (0.357)	0.943*** (0.258)
Obs.	5086	5086	4904	4904
R-squared	0.061	0.053	0.072	0.078

*** p<0.01, ** p<0.05, * p<0.1

2. 替换因变量

本文在实证论证过程中，使用了 $NCSKEW_{i,t+1}$ 和 $DUVOL_{i,t+1}$ 两种衡量股价崩盘的指标。使用两种指标获得的实证结果均一致。因此本文的结论具有一定的稳健性。

3. 替换自变量

本文使用了积极词汇、消极词汇词频的自然对数代替净正向语气进行稳健性检验。具体的，使用公式（4）中的 POS 和 NEG 的自然对数，分别记为 $\ln POS_{i,t}$ 和 $\ln NEG_{i,t}$ ，并替换自变量 $TONE_{i,t}$ 进行上述的多元回归分析。 $TONE_{i,t}$ 与 $\ln POS_{i,t}$ 、 $\ln NEG_{i,t}$ 的皮尔森相关性系数分别为-0.0513和-0.5948，相关性系数均在1%水平上显著。表7展示了替换自变量之后的回归结果。列（1）（2）重复表4的多元回归， $\ln POS_{i,t}$ 的系数分别为-0.067和-0.116， $\ln NEG_{i,t}$ 的系数分别为0.153和0.153， $\ln NEG_{i,t}$ 的系数均在1%水平上显著。根据公式， $\ln NEG_{i,t}$ 与 $TONE_{i,t}$ 是负相关关系，因此 $\ln NEG_{i,t}$ 系数为正印证了年报文本语气与股价崩盘的负相关关系。列（3）（4）（5）（6）重复表5的多元回归，其中列（3）和（4）是“正向组”的回归，（5）和（6）为“负向组”的回归。比较 $\ln POS_{i,t}$ 和 $\ln NEG_{i,t}$ 的系数可以得知，“负向组”的 $\ln POS_{i,t}$ 和 $\ln NEG_{i,t}$ 系数均大于正向组，且 $\ln NEG_{i,t}$ 系数的差异要远大于 $\ln POS_{i,t}$ 系数的差异。使用替换后的自变量重复进行上述的回归，所得的结果仍能支持本文的结论。此外根据列（1）（2）（5）（6）的结果可知， $\ln NEG_{i,t}$ 系数的绝对值均要大于 $\ln POS_{i,t}$ 系数的绝对值，说明在本文的结论中，年报的消极词汇对股价崩盘的影响程度要高于积极词汇对年报文本的影响程度。这也从侧面印证了与提供积极语气，提振投资者信心相比，年报文本语气对股价崩盘的影响更多地是通过传递更多的消极语气，激发投资者避险情绪而进行的。

表7：年报文本语气词词频与股价崩盘指标回归结果

	全样本		$ABTONE_{i,t} \geq 0$		$ABTONE_{i,t} < 0$	
	(1)	(2)	(3)	(4)	(5)	(6)
	$NCSKEW_{i,t+1}$	$DUVOL_{i,t+1}$	$NCSKEW_{i,t+1}$	$DUVOL_{i,t+1}$	$NCSKEW_{i,t+1}$	$DUVOL_{i,t+1}$
$\ln POS_{i,t}$	-0.067 (0.047)	-0.116*** (0.038)	0.106 (0.102)	0.006 (0.065)	-0.148 (0.091)	-0.216*** (0.065)
$\ln NEG_{i,t}$	0.153*** (0.049)	0.153*** (0.036)	0.006 (0.089)	0.040 (0.054)	0.165 (0.104)	0.229*** (0.075)
(控制变量)	控制	控制	控制	控制	控制	控制
(行业)	控制	控制	控制	控制	控制	控制
(年份)	控制	控制	控制	控制	控制	控制
常数项	0.155 (0.329)	0.337 (0.217)	0.178 (0.365)	0.249 (0.243)	0.290 (0.438)	0.473 (0.298)
Obs.	11483	11483	5906	5906	5577	5577

R-squared	0.055	0.054	0.051	0.046	0.067	0.071
	*** p<0.01, ** p<0.05, * p<0.1					

五、结论、启示与不足

本文基于 2010 年至 2018 年 A 股上市公司年报的数据，利用 python jieba 库分词函数对文本的非定量信息进行分析，研究年报文本语气对公司股价崩盘的影响。结果表明，年报文本的语气能够影响公司的股价崩盘风险，年报文本净正向语气越少，则股价崩盘风险越大。这种影响关系与管理层对年报文本语气操纵情况有所联系，当年报文本存在负向的可操纵语气时，年报文本语气与股价崩盘风险显著负相关；而当年报文本存在正向可操纵语气时，年报文本语气对股价崩盘风险的影响不显著。这些证据表明，年报文本语气信息能够提供增量信息，影响投资者的投资决策，而且这种影响主要通过消极的语气引起投资者避险情绪，进而增大股价崩盘风险的方式进行的。

随着互联网与计算机技术突飞猛进地发展，大数据与金融科技为投资者们提供了更多更为深刻的数据挖掘技术。如今的投资者能比以往从同一份文件中获取更多的信息，这对监管部门提出了更高的要求。本文的研究具有一定的政策启示：首先，年报文本的语气是有价值的计量指标，监管部门应当积极利用计算机文本挖掘技术，从非定量信息的角度加强对上市公司信息披露质量的评估与监管。其次，年报文本语气与会计信息一样存在可操纵空间，监管部门应当进一步规范公司非定量信息内容的披露情况，完善相关的制度与环境建设。最后，年报文本语气可以作为衡量股价崩盘风险的先行指标，监管部门可以将其作为非定量信息因子纳入对总体金融风险评估的模型之中，并依此开展更多关于行为金融的研究。

本文的研究仍然存在着不足：第一，虽然本文发现了积极语气与消极语气分别对股价崩盘的影响有所不同，但没有对其传导机制的不同进行更加深入的分析。第二，本文仅研究了年报文本对一年内股价崩盘的影响，但实际中文本语气的背离可能持续多年后一并爆发，本文未采用时间序列分析进行更深入的研究。这些方面是本文的不足，也是未来更为深入研究的方向。

附 录

附表1：主要变量皮尔森相关性系数

	$NCSKEW_{i,t+1}$	$DUVOL_{i,t+1}$	$TONE_{i,t}$	$ABTONE_{i,t}$	$ABTONE_FE_{i,t}$	$RET_{i,t}$	$SIGMA_{i,t}$	$TURNOVER_{i,t}$	$ROA_{i,t}$	$LEV_{i,t}$	$BM_{i,t}$	$SIZE_{i,t}$	$HOLD_{i,t}$	$ACCM_{i,t}$
$NCSKEW_{i,t+1}$	1.000													
$DUVOL_{i,t+1}$	0.872***	1.000												
$TONE_{i,t}$	-0.046***	-0.072***	1.000											
$ABTONE_{i,t}$	-0.046***	-0.065***	0.777***	1.000										
$ABTONE_FE_{i,t}$	-0.045***	-0.064***	0.776***	0.999***	1.000									
$RET_{i,t}$	-0.069***	-0.103***	-0.027***	0.002	0.002	1.000								
$SIGMA_{i,t}$	-0.075***	-0.094***	-0.053***	0.015	0.016*	0.658***	1.000							
$TURNOVER_{i,t}$	-0.053***	-0.044***	0.050***	-0.010	-0.010	-0.123***	-0.467***	1.000						
$ROA_{i,t}$	0.022**	0.006	0.146***	0.014	0.010	0.066***	-0.089***	-0.014	1.000					
$LEV_{i,t}$	-0.019*	-0.020**	0.048***	-0.028***	-0.026***	-0.020**	-0.051***	0.042***	-0.343***	1.000				
$BM_{i,t}$	-0.065***	-0.045***	0.057***	-0.039***	-0.039***	-0.324***	-0.351***	0.150***	-0.244***	0.500***	1.000			
$SIZE_{i,t}$	-0.055***	-0.062***	0.012	-0.068***	-0.068***	-0.040***	-0.143***	0.014	0.020**	0.471***	0.624***	1.000		
$HOLD_{i,t}$	-0.045***	-0.049***	0.001	-0.086***	-0.087***	-0.003	-0.037***	0.026**	0.072***	0.118***	0.173***	0.228***	1.000	
$ACCM_{i,t}$	0.009	-0.004	0.064***	0.015	0.016*	-0.051***	-0.034***	0.010	0.147***	0.005	0.089***	0.100***	0.006	1.000