专题报告——金融期货

基于文本挖掘的商品期货投资策略探讨

欧阳静宜　分析师（金融工程）

从业资格号：F3020984

Tel:　8621-63325888-4268

Email:　jingyi.ouyang@orientfutures.com

报告日期：　2017 年 8 月 17 日

★大数据金融产品现状：

目前有 24 只公募基金发行了基于大数据的基金，基金规模约 150 亿。绝大部分基金都借助了互联网行业的数据优势，采取数据合作的方式。主要的数据有三类：搜索引擎的用户搜索数据、财经门户网站的股票关注度、电商的销售数据。大数据已经在股票市场中有所应用，在期货市场中的应用较少。期货市场标的少，热点更加集中的特点，使得文本分析有广阔的空间。

★文本数据库、词库搭建：

文本分析的基础是数据库的扩充，正在逐步搭建中，目前在黑色系的新闻方面已经有较为全面的采集。分词词库在原有的词库基础上，综合整理添加了搜狗金融词库和 wind 上市公司信息，并且补充了期现货市场专用术语，能够满足证券期货市场文本分词需求。

★应用一：热点跟踪

改良了依赖于文本数量的词频监测，改用词频排名，有效剔除“钢铁”、“产量”等频常用词汇，精准定位“地条钢”、“天气”等期市热点和炒作周期。

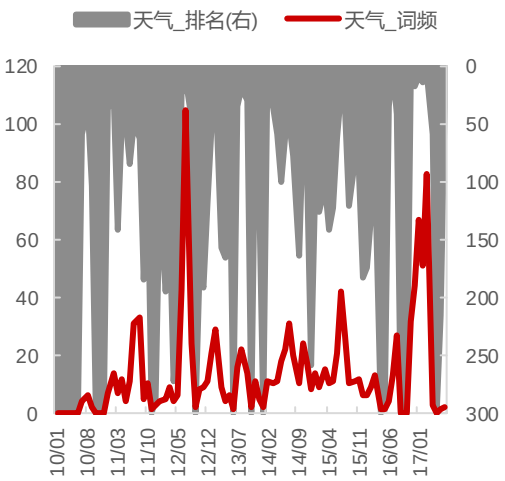
★应用二：舆情监控

采用扩充情感词库的方法，提取文本中的关键词，直接做出对品种、供需和方向上的判断，并综合整理成情绪指标，方向上与黑色系品种价格走势一致。基于舆情指标的交易策略，胜率 51.67%，年化收益率 23.84%（1 倍杠杆），基于文本的交易策略还有很大的优化空间。

★应用展望：

本篇分析是基于情感词库来识别文本中的价格影响，文本的表达方式和词库的完备程度会对结果产生较大的影响，这是一个局限。另一种基于机器学习的方法，需要大量有情感标注的语料库，如有评级的分析师研究报告。尽管现阶段机器学习对于本文分析的新闻素材适用性不高，未来的潜力无限。

“天气”炒作周期



舆情指数及价格走势

目录

图表目录

1、“金融+大数据”行业现状

这是一个信息爆炸的年代。获取信息的渠道空前的广泛，网络平台取代了传统纸媒，成为主流的信息传播方式，同时也大大地降低了传播的成本，新闻门户网站百花齐放。不仅如此，话语权也不仅仅掌握在权威的媒体的手中，每个人都有在社交网络上表达观点的途径。所以，每时每刻都有海量信息铺天盖地而来。

这也是一个信息匮乏的年代。因为人总有盲点，于是就有人雇佣无人机侦探港口实时的吞吐量，并且，人的精力是有限的，在海量的信息面前，难以保持对信息的敏感度。尤其是在金融市场中，有效市场假说告诉我们，价格反映了信息，如果我们能先知一步，或是对市场动态更加敏感，那我们就是站在更高的位置。

金融+大数据的概念并不陌生，早已有相关的基金产品发行，目前据不完全统计目前仍在存续期的大数据基金共有 24 只，规模总计近 150 亿，其中，南方大数据 100 发行最早（2015 年 4 月 24 日），规模最大（57.65 亿）。基于大数据的特征，这些基金大部分选择与互联网行业合作的形式，互联网公司提供数据，基金公司基于这些数据做出交易决策。根据数据的类型，大致可以分为以下 4 种形式：

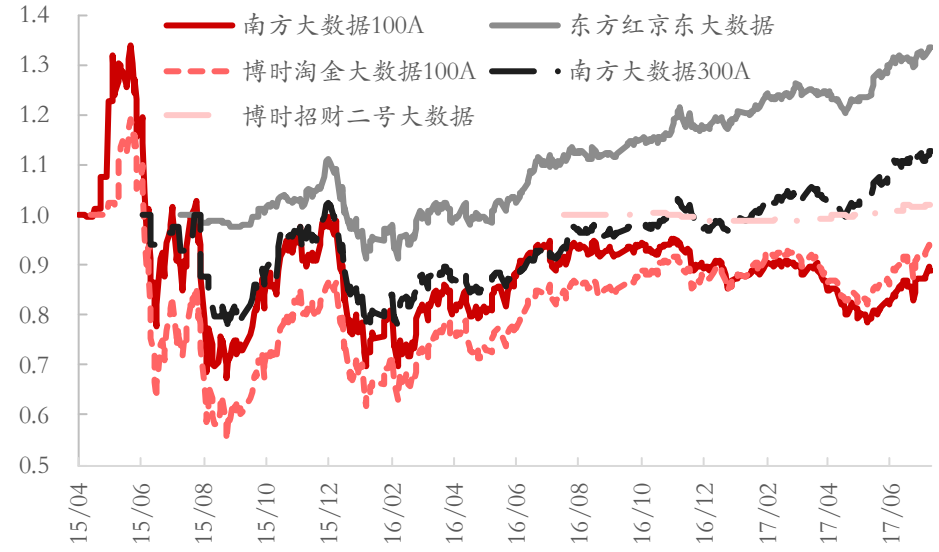
第一种：与百度和奇虎 360 等搜索引擎公司合作，获取用户的搜索热点。这类基金总体规模约 9 亿，整体规模较小，代表基金有广发百发大数据策略成长 A，它是与百度合作。

第二种：与东方财富网等财经类门户网站合作，获取用户的股票关注、访问量，在第一种的基础上更有针对性。这类基金的规模约 76 亿，是四种中最大的，代表基金为南方大数据 100A，它是与新浪财经合作。

第三种：与电商或银联合作，获取消费者的消费信息、点击量、浏览量、收藏量及客户评价。这种形式适合发行消费行业主题基金，整体规模约为 58 亿，代表基金有：东方红京东大数据和博时淘金大数据 100A，它们分别与京东和蚂蚁金服（淘宝）合作。

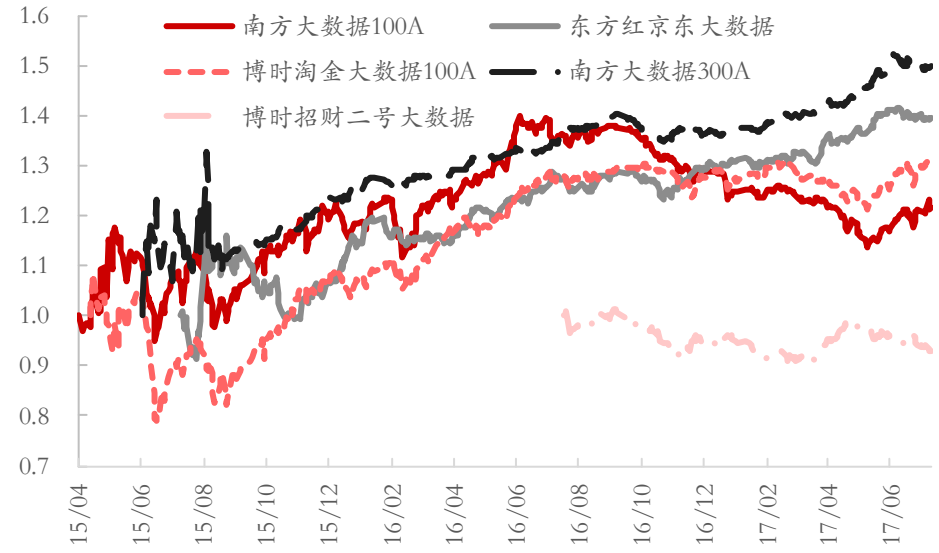
第四种：自主研发的基金，他不与单一渠道的合作方合作，而是自主搜集网络信息数据作为交易决策基础，这类目前仅有 2 只公募基金，总体规模约 4.5 亿，代表基金为浙商大数据智选消费。

图表 1：大数据基金净值走势



资料来源：Wind，东证衍生品研究院

图表 2：大数据基金相对上证综指走势



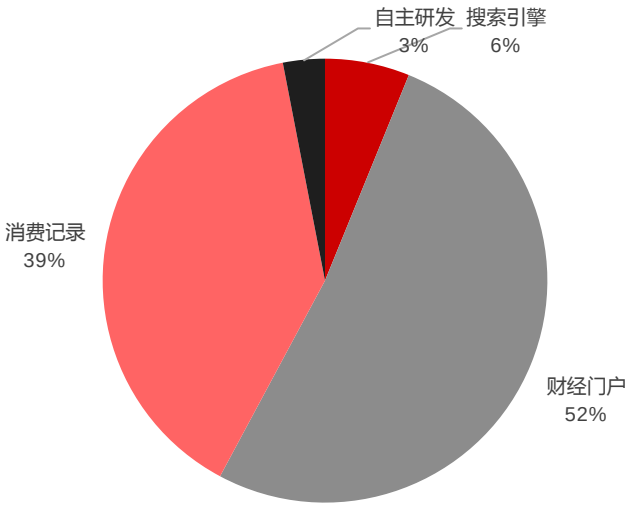
资料来源：Wind，东证衍生品研究院

图表 3：大数据基金产品（截至 2017/07/31）

名称	成立日期	基金规模（亿）	合作方	数据
南方大数据 100A	2015/4/24	57.64	新浪	股票页面访问热度
东方红京东大数据	2015/7/31	30.56	京东	京东大数据（包括京东电商的销量、浏览量、点击量、客户评价、客户收藏量等基础数据）
博时淘金大数据 100A	2015/5/4	13.81	蚂蚁金服	网络交易数据
南方大数据 300A	2015/6/24	9.05	新浪	股票页面访问热度
博时招财二号大数据	2016/8/9	8.07	蚂蚁金服	网络交易数据
嘉实腾讯自选股大数据	2015/12/7	5.50	腾讯	腾讯自选股大数据
浙商大数据智选消费	2017/1/11	3.62	自主研发	包括但不限于支付、电商、论坛、新闻、舆情、市场等数据
博时淘金大数据 100I	2015/5/4	2.76	蚂蚁金服	网络交易数据
广发百发大数据策略成长 A	2015/11/18	2.73	百度	用户搜索行为
博时银智大数据 100	2016/5/20	2.31	银联	基于中国银联交易流水的消费类行业
广发百发大数据策略价值 A	2017/6/16	1.94	百度	用户搜索行为
泰达宏利同顺大数据 A	2016/2/23	1.83	同花顺	互联网新闻数据、新闻点击数据
大成互联网+大数据 A	2016/2/3	1.41	奇虎 360	用户搜索行为
广发百发大数据策略价值 E	2017/6/16	1.22	百度	用户搜索行为
南方大数据 300C	2015/6/24	1.08	新浪	股票页面访问热度
广发百发大数据 E	2015/9/14	0.87	百度	用户搜索行为
银华大数据	2016/4/7	0.74	自主研发	使用大数据分析技术，挖掘出有价值的证券
海富通东财大数据	2016/1/29	0.42	东方财富	东方财富大数据（包括股票关注度、点击量等投资者行为数据）
广发百发大数据 A	2015/9/14	0.42	百度	用户搜索行为
招商财经大数据策略	2016/11/2	0.41	东方财富	投资者行为数据因子，包括股票关注度、点击量等投资者行为数据
广发百发大数据策略成长 E	2015/11/18	0.40	百度	用户搜索行为
大成互联网+大数据 C	2017/1/18	0.06	奇虎 360	用户搜索行为
泰达宏利同顺大数据 C	2017/2/9	0.03	同花顺	互联网新闻数据、新闻点击数据
南方大数据 100C	2017/2/23	0.01	新浪	股票页面访问热度
总计		146.98		

资料来源：Wind，东证衍生品研究院

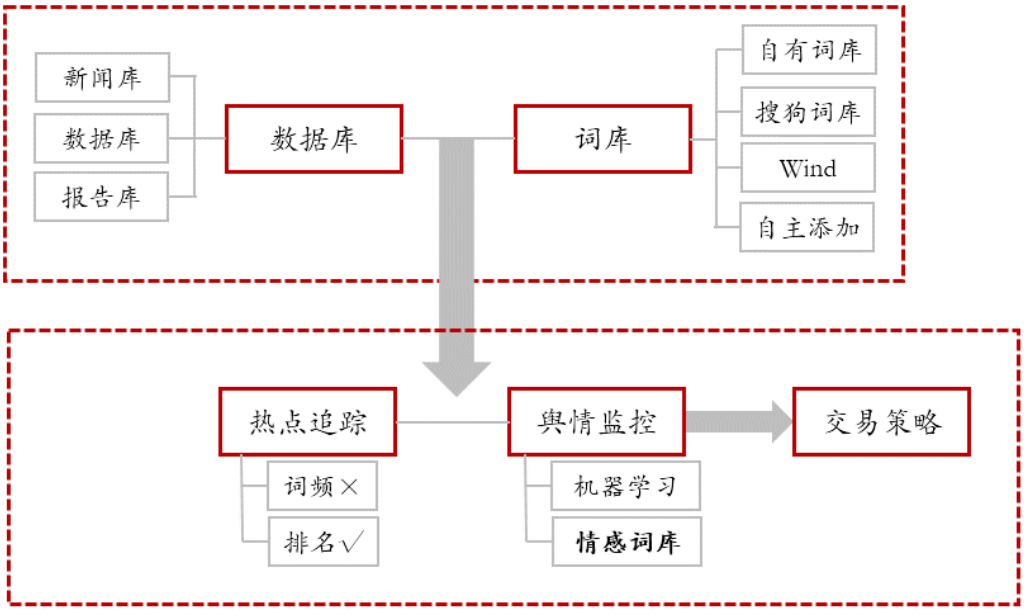
图表 4：大数据基金规模构成



资料来源：Wind，东证衍生品研究院

文本分析在国内期货市场的运用还很少。期货市场天然有应用文本分析的优势，其一，大宗商品在国民生产生活中的基础地位决定了文本量丰富，其二，期货市场活跃的标的数量有限，这易于将文本与标的直接挂钩。

图表 5：分析结构



资料来源：东证衍生品研究院

2、分词词库建立

分词是文本分析的基础，中文分词方面，目前 R 和 python 已经有十分成熟的 jieba 分词开源程序包，可以实现基本的分词、词性标注、关键词提取等功能。下面以 R 语言处理结果为例：

图表 6：分词示例

文本	海外铜矿供给干扰因素解除。					
分词	海外	铜矿	供给	干扰	因素	解除
词性标注	s	n	vn	v	n	v
关键词打分	铜矿		干扰		解除	
	8.38725		7.31865		7.29133	

资料来源：东证衍生品研究院

jiebaR 包中已经内涵了人民日报等语料库中提取的词汇，能够从容应对语法正确的日常文本。但在专业领域内还是有所欠缺，不一定能识别出专业词汇，尽管这一点能够通过长期的训练得到改善，但是直接利用 jiebaR 的词库导入功能更加便捷。

我们整理了以下 3 个词汇来源作为词库扩充：

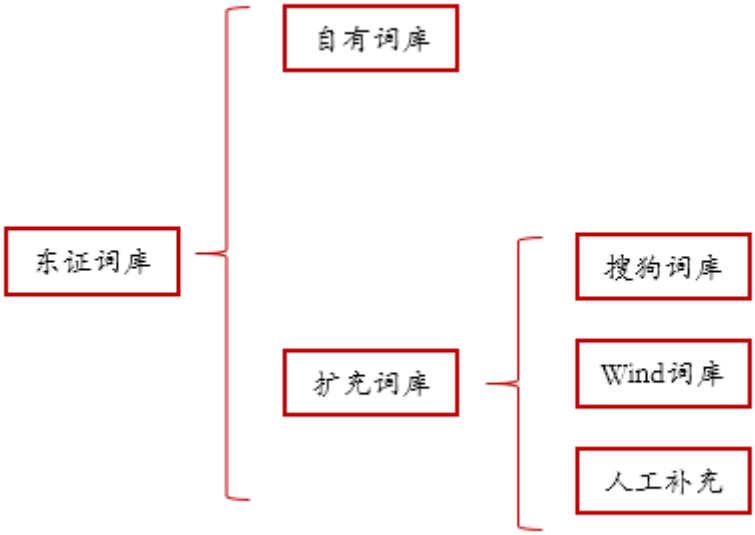
搜狗词库：搜狗拼音内置的扩充词库，词汇多且广，但是整体质量不高，稍欠专业，按需求纳入金融词库、股票基金词库、商品期货词库和商品产业链词库。词汇包括：安全度、保证金账户、波段操作、银行存款利率、远期汇率。下载的词库为加密格式，可以用 R 方便地转化为 TXT 格式使用。

Wind 词库：全面覆盖了股票和基金市场，专业名词最为准确，股票市场包括全部上市公司的名称、代码、简称、高管及前任高管名称、公司网址等。基金词汇包括基金名称、基金管理人、托管人等。词汇包括：平安银行、平安银行股份有限公司、谢永林此类。而期货市场的词汇十分欠缺。全部数据用 wind 插件导出。

人工补充的词汇：期货市场相对小众，搜狗词库和 wind 都没有完全覆盖，所以着重添加了期货词汇，除此之外，还有从实践中总结出来的遗漏词汇如软商品、黑色系、去产能、供给侧改革等。

整合以上词库，去除重复词汇，共计扩充 45052 个词汇（截至 2017 年 7 月 26 日）。

图表 7：词库扩充



资料来源：东证衍生品研究院

以这句话为例，展示分词效果：“上期所巨量天胶库存预计会对 9 月合约期价构成持续压制，并在 11 月合约交割流出后也难被现货市场有效消化，届时或给国内天胶供给端造成明显压力。”

图表 8：分词示例（词库扩充前）

[1]	"上期"	"所"	"巨量"	"天胶"	"库存"	"预计"	"会"	"对"	"9"	"月"	"合约"
[12]	"期价"	"构成"	"持续"	"压制"	"并"	"在"	"11"	"月"	"合约"	"交割"	"流出"
[23]	"后"	"也"	"难"	"被"	"现货"	"市场"	"有效"	"消化"	"届时"	"或"	"给"
[34]	"国内"	"天胶"	"供给"	"端"	"造成"	"明显"	"压力"				

资料来源：东证衍生品研究院

图表 9：分词示例（词库扩充后）

[1]	"上期所"	"巨里"	"天胶"	"库存"	"预计"	"会"	"对"
[8]	"9"	"月"	"合约"	"期价"	"构成"	"持续"	"压制"
[15]	"并"	"在"	"11"	"月"	"合约"	"交割"	"流出"
[22]	"后"	"也"	"难"	"被"	"现货市场"	"有效"	"消化"
[29]	"届时"	"或"	"给"	"国内"	"天胶"	"供给端"	"造成"
[36]	"明显"	"压力"					

资料来源：东证衍生品研究院

除了在词库中加入专业名词等重要词汇，而且为了便于分析，也要将出现频率很高但实际意义不重要的词纳入停止词库，如：的、仍、与、个、年、超等，共计约 904 个。但停止词在某些情感分析中也有重要作用，在实际分析过程中酌情使用。

图表 10：分词示例（加入停止词后）

```
 [1] "上期所"   "巨量"     "天胶"   "库存"   "预计"   "会"     "合约"
 [8] "期价"     "持续"     "压制"   "11"     "合约"   "交割"   "流出"
[15] "难"       "现货市场" "消化"   "届时"   "国内"   "天胶"   "供给端"
[22] "压力"
```

资料来源：东证衍生品研究院

在经过词库扩充后，重要的专业词汇能够有效的识别出来，在经过停止词的过滤，使得文本在保有原意的基础上，去除了对文本意义影响较少的词，使得最终呈现出来的分词结果更加精炼、准确。

目前，我们已收录进入关键词词库的数量已有 4.5 万个，停止词库近 1 千，但是这依然是不够的，随着行业的发展和世界的进步，词库也要不断丰富升级。

3、热点监控

在文本热点监控上，商品市场有一定的优势。其一，商品期货市场活跃的标的较少，相关文本更容易对应到标的。其二，大宗商品的在生产生活中的基础地位决定了新闻数量多，而且与国际市场联动，新闻来源广泛，少数品种面对着大量的文本信息，这就更需要文本分析批量处理。

商品市场主要的新闻门户网站主要包括以下：

图表 11：商品期货主要信息门户

综合	工业品	农产品	黑色系	有色金属
Bloomberg	轮胎世界网	美国农业部	我的钢铁网	中国冶金报
Reuters	中国橡胶贸易信息网	天下粮仓	中国煤炭经济研究会	镍吧
国家统计局	卓创资讯	中华粮网	西本新干线	上海有色网
海关总署	隆众石化	广西糖网	中国煤炭资源网	
华尔街见闻	化工在线	中国棉花信息网	秦皇岛煤炭网	
期货日报	中纤网	中国棉花网		
	中国橡胶网	饲料行业信息网		
		农产品期货网		
		中国玉米信息网		

资料来源：东证衍生品研究院

注：排名不分先后

目前，我们新闻数据库的建设还在完善中，目前涵盖了期货日报、我的钢铁网、广西糖网、中国橡胶网和镍吧的所有新闻，共计超 120 万条信息。将来我们的数据库会不断的完善上述新闻数据，但因此本篇报告会以钢铁产业为例。基于上述新闻数据，对新闻标题做分词处理，并统计词频，并绘制出新闻图，如下：

图表 12：2017 年商品市场新闻高频词

资料来源：东证衍生品研究院

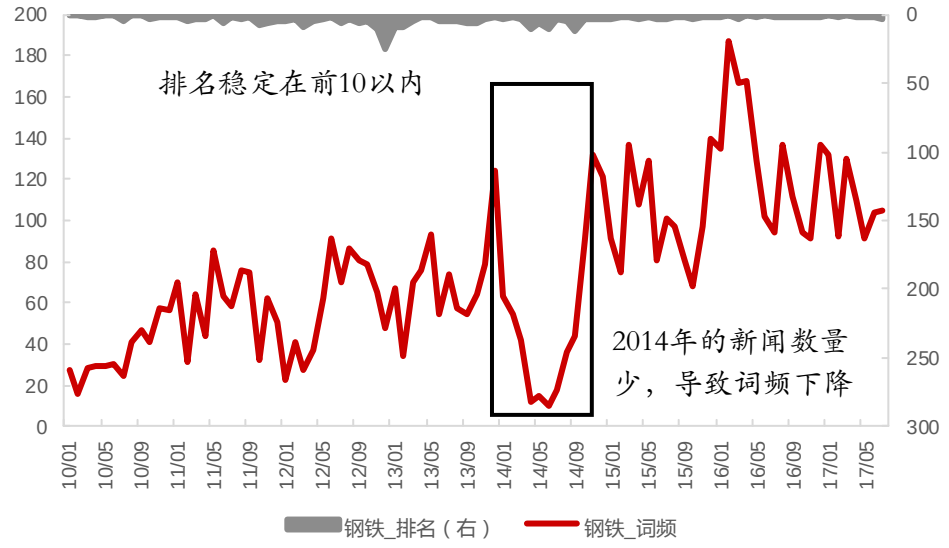
图表 13：2015 年商品市场新闻高频词

资料来源：东证衍生品研究院

可见在 2017 年，新闻谈论的更多的是煤炭、钢铁、黄金、原油、地条钢、产量、减产与过剩。作为对比，在 2015 年新闻报道的关键词是十分相似的，也是谈论煤炭、原油、产量，但实际上这两年的市场情况差别很大。2015 年商品市场还处于熊市当中，在年底 11 月份的时候政府提出了“供给侧改革”，商品市场的牛市由此开启，此后市场上的炒作热点围绕着。年年岁岁花相似，岁岁年年人不同，按词频来提取关键词黄金、原油、钢铁等商品名称永远将会是市场的热点，但实际上市场关注的焦点是会转移的。所以，我们认为只是单纯由词频来提取关键词在实际操作上是缺乏指导意义的。

由此，我们提出按词频排名来追踪市场热点。作为对比，我们选取了“钢铁”和“地条钢”两个词语的词频与词频排名。“钢铁”是商品期货市场中具有代表性的常用词汇，单从词频来看，呈现出逐年升高且波动十分剧烈的特征，原因在于市场中的信息数量趋势一定是增长的，所以大部分词语的频率都会抬升，而且大宗商品在工业生产中的基础地位决定了不会像 BP 机一样轻易消失在视野中。但也看到了在 2014 期间词频有大幅的下滑，原因在于词频依赖于文本数量，当年的新闻数量极少，这一点可以通过增加信息来源途径来缓解，但这也揭示了**按词频分析的被动性：词频的下降可能不是由于热度降温，而是由于市场中的声音少了。**从排名来看，“钢铁”在近 7 年都是热门词汇，排名稳定在前十，即使是在词频大幅下降的 2014 年同样如此。

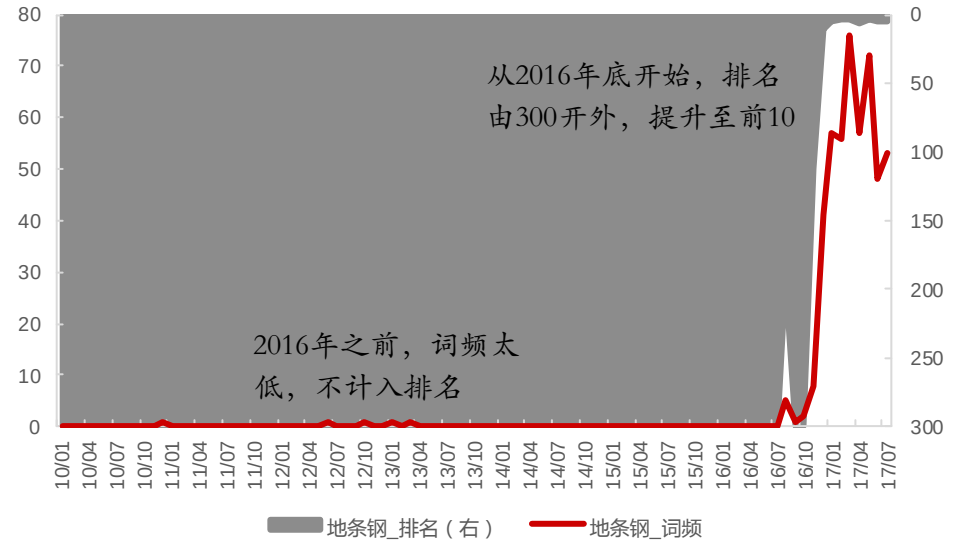
图表 14：“钢铁”词频及排名



资料来源：东证衍生品研究院

再看“地条钢”，它是典型的热点词汇。它于 2010 年 12 月在新闻中首次报道，但并没有热度，排名一直在 300 名之外。直达 2016 年 8 月出现了炒作的苗头，并于同年 12 月份开始被大量报道，排名进入前 5 。“地条钢”的词频和排名都呈现出相似的形态，都在炒作时期大幅上升。

图表 15:“地条钢”词频及排名

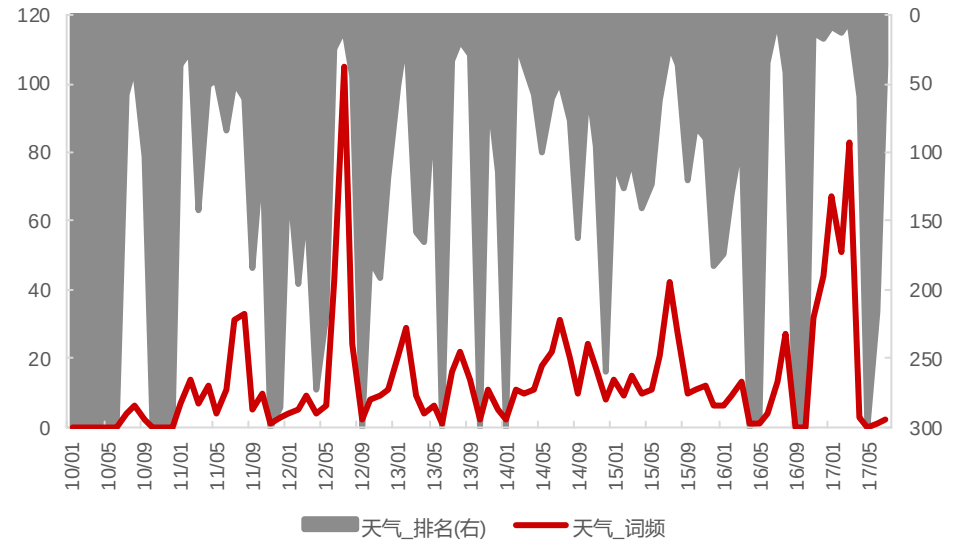


资料来源：东证衍生品研究院

“天气”一词有明显的周期性特征。这是农产品特有的“天气”炒作，在每年 7、8 月份往往有明显的排名上升的行为，尽管市场中可能对天气炒作会形成一定的预期，但每年的情绪强弱都有所不同，而且在 1-3 月份也有炒作的迹象，这种反常行为更值得投资者关注。

综合以上分析，市场中的热点不在于词频的多少，因为这依赖于文本数量，也不在于排名靠前，因为基础词汇使用量排名一直很高，而在于排名的上升幅度。

图表 16:“天气”词频及排名



资料来源：东证衍生品研究院

基于以上观点，以黑色系商品新闻为例，筛选出 2016 年 8 月至 2017 年 7 月以内，排名曾在 1 个月内上涨 100 名的词汇，并以当前词频绘制新闻词云。可以看出钢铁、煤炭、美国和中国这种基础词汇从图中消失了。

图表 17：2017 年商品市场新闻关键词

资料来源：东证衍生品研究院

4、舆情监控

在新闻数据库有一定的积累后，除了实时跟踪市场热点以外，还可以监控市场中的舆情。

在本文第一部分揭示了目前大数据公募基金的数据基础，那些基于用户搜索行为、股票关注度、商品销量的数据分析，归根结底依然还是纯数字的研究方式。目前 A 股市场不能做空，股票的热度的确能反映市场中的看多情绪，但对于一个可以做空的期货市场，这就不一定了。挖掘市场对某类品种的舆情，不仅要从量上着手，也要深入文本的情感中去。

文本舆情监控主要有两种思路：

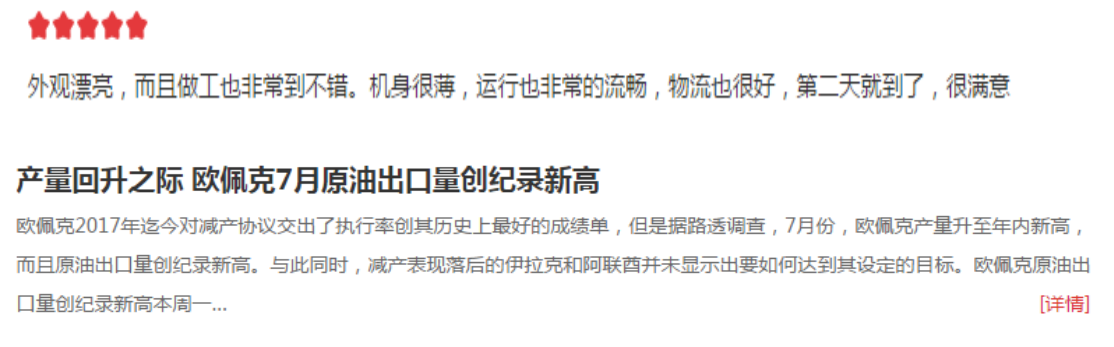
第一种：基于机器学习的方法。这种方式需要向机器提供大规模有人工标注的语料库作为训练集，机器再提取这些信息的文本特征，构建分类器来实现情感的分类。如豆瓣电影和商品的评论附带了打分信息，并且情绪表达直接，这种语料库天然适合机器学习的方法。

第二种：基于情感词典的方法。这种方法是需要大量已标注倾向的情感词典，如“乐观”、“看涨”对应的是正面的情绪，“有限”、“堪忧”等对应的是负面的情绪，它们分别纳入正面词库和负面词库，除此之外，还需构建否定词库和程度词库，诸如此类。

这种舆情监控技术的可应用范围十分广泛，比如，电商产品评价、电影打分和社交网络舆情监控。在本文的第一部分也提到了，某些大数据基金确实使用了类似的对电商商品或股票评论信息。以上评论的文本信息情绪表达更加直接，易于判断情感的倾向，不仅如此这种文本的往往自带了情感标注，可以直接作为机器学习的语料库，但是这类文本往往是用口语化、非规范化的表达方式，甚至使用了网络用语，所以情感词典难以覆盖。

我们数据库目前的文本内容是以新闻为主，新闻素材自身不带有情感标注，且整体数量庞大（目前约 120 万条），用人工标注的方法工作量将十分繁重。另外，新闻立场中性，对标的价格的影响需要分析其内在传导逻辑，而不是情感偏向，如漂亮、不错、流畅等词语，即便是人工标注也容易产生误判。新闻文本的优点在于表达上更加的规范，这就使得情感词典更易于覆盖。所以，机器学习的方法暂且按下不表，本篇研究我们主要着力于构建情感词典的方法。

图表 18：两类表达方式对比



资料来源：东证衍生品研究院

首先，构建情感词典。目前，已经有搭建完备的成熟词库可以使用，如知网和台湾大学都有整理，但是这种通用词库在证券期货领域的适用性较差，我们依然选择自主构建。主要分为五个子库：

- 标的词库用于提取文本对应的标的。
- 升降词库用于提取文本中涨跌升降等信息，仅描述数量升降，不是对价格的最终影响方向。
- 供需词库用于判断文本描述的是供给面还是需求面的信息。因为供给面和需求面的涨跌对大宗商品价格影响的方向是截然不同的。
- 否定词库用于提取文本中的否定词。
- 特殊词库除了供需面的涨跌新闻会对商品价格构成影响外，还有“环保”、“罢工”、“监管”等新闻也需要识别出来。

图表 19：舆情分析示例

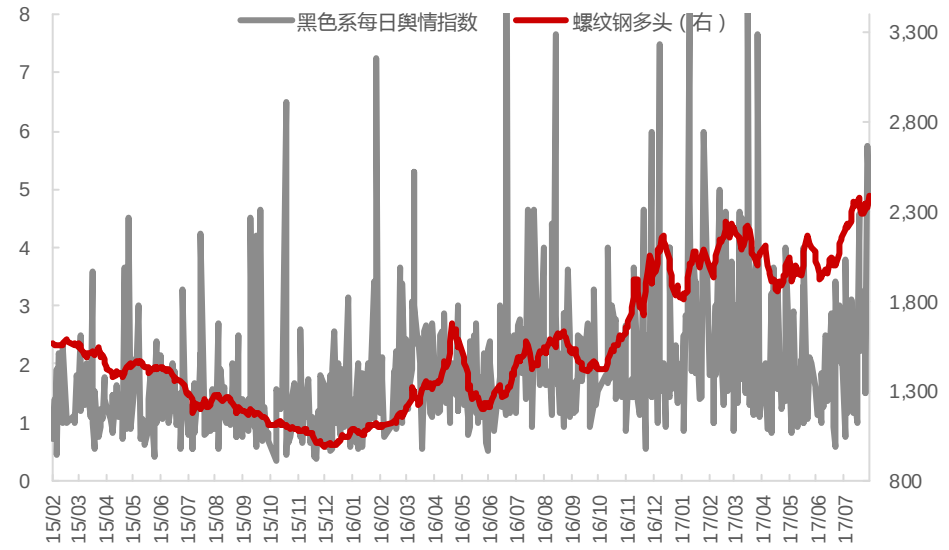
	文本	品种	供需	升降	影响正负
新闻一	河北钢铁集团淘汰落后产能 400 万吨	钢铁	供给	降	正
新闻二	七大煤企与中煤重组的煤矿实行省级安全监管	煤炭	供给	降	正
新闻三	8 月意大利新车销量下降 20.23%	橡胶	需求	降	负

资料来源：东证衍生品研究院

将所有的新闻信息都与词典相匹配，计算每条新闻的正负面影响。$num.positive_t$ 和 $num.negative_t$ 分别为 t 交易日当日的正、负面影响新闻数量，非交易日的新闻计入下一交易日。每日的舆情监控指标构建如下：

$$sentiment_t = \frac{num.positive_t}{num.negative_t}$$

图表 20：黑色系每日舆情指数



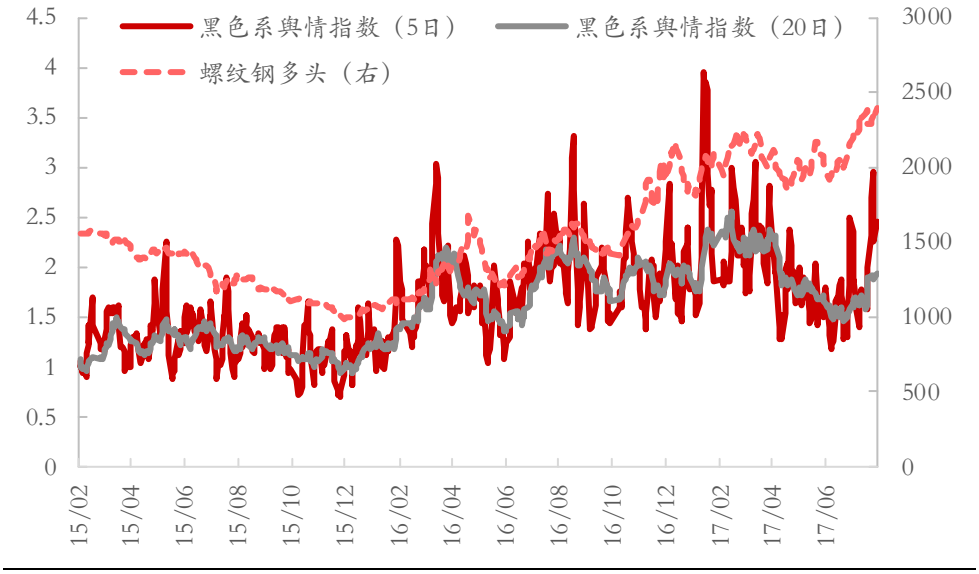
资料来源：东证衍生品研究院

随着新闻库的扩大，舆情指数会变得更平滑。为了对投资更有指导意义，还构建了更为平滑的移动 5 日和移动 20 指标。

$$sentiment5_t = \frac{\sum_{i=t-6}^{t-1} num.positive_i}{\sum_{i=t-6}^{t-1} num.negative_i}$$

$$sentiment20_t = \frac{\sum_{i=t-21}^{t-1} num.positive_i}{\sum_{i=t-21}^{t-1} num.negative_i}$$

图表 21：黑色系多日舆情指数

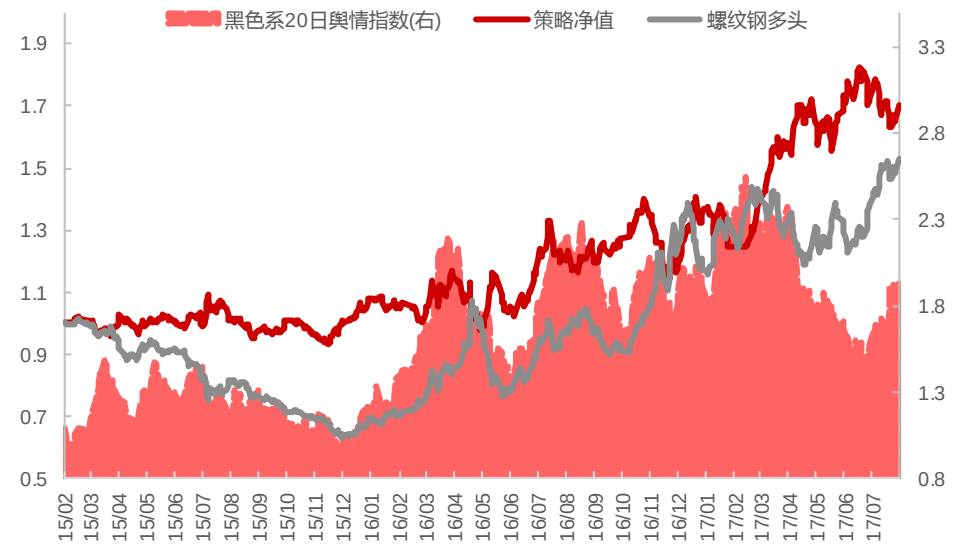


资料来源：东证衍生品研究院

5、策略回测

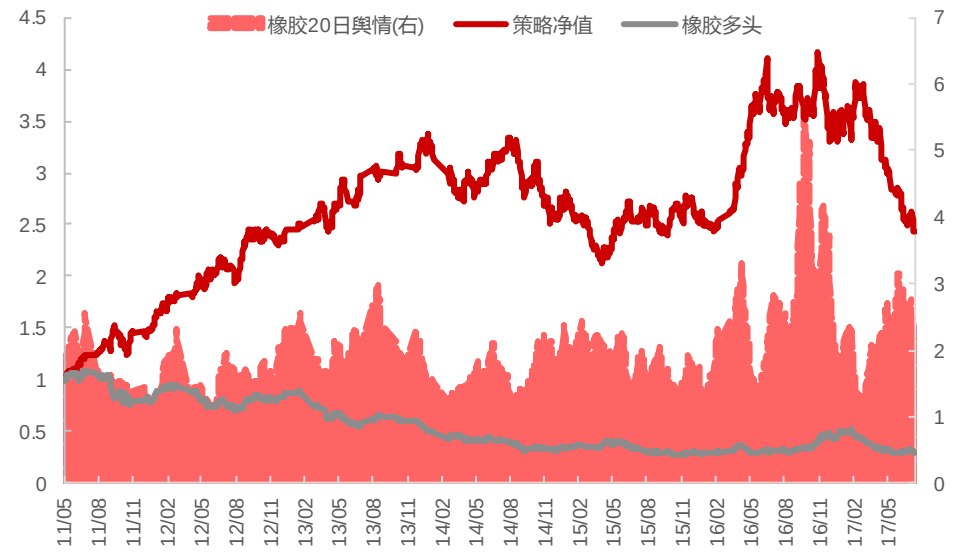
构建一个基于舆情监控的简单交易策略：当 5 日指数高于 20 日指数时，持有标的主力多头；当 5 日指数低于 20 日指数时，持有标的主力空头。往往商品价格上涨之后，随后会有相关的新闻报道，为避免未来信息影响当前决策，20 日舆情指数和 5 日舆情指数计算，如上述计算公式所列示，均不包括当日的新闻信息。成交价以结算价计，开平仓交易成本计 0.02%。当开仓亏损 5%以上，平仓止损，并且下次不以同一方向开仓。

图表 22：钢铁情绪指标及策略回测



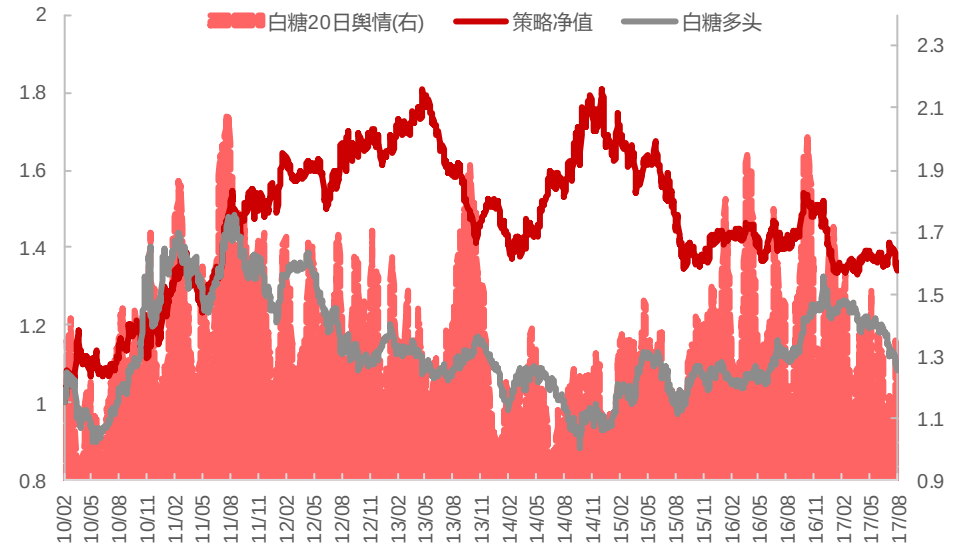
资料来源：东证衍生品研究院

图表 23：橡胶情绪指标及策略回测



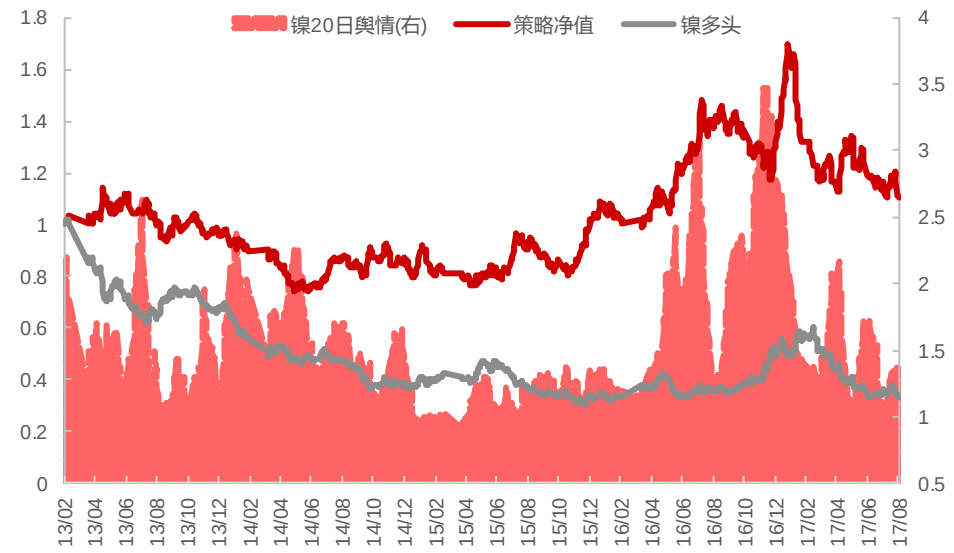
资料来源：东证衍生品研究院

图表 24：白糖情绪指标及策略回测



资料来源：东证衍生品研究院

图表 25：镍情绪指标及策略回测



资料来源：东证衍生品研究院

图表 26：回测结果（1 倍杠杆）

	黑色	橡胶	白糖	镍
年化收益率	23.84%	15.40%	4.16%	2.23%
年化波动率	24.69%	24.39%	14.18%	23.97%
夏普比率	0.97	0.63	0.29	0.09
胜率	51.88%	52.58%	50.88%	50.65%

资料来源：东证衍生品研究院

基于这种文本分析的策略，词库的建完备性对策略收益的影响是最大的，我们也在不断的完善中。策略目前波动率较大，夏普比率普遍较低。在上述回测中，除了 5%的止损条件以外，无其他参数优化，风险较大也在情理之中，未来还有很大的深入探究的空间，回测结果仅做投资参考。

6、总结与展望

文本分析当前在各个领域都是一个炙手可热的话题，这个技术在正在也被广泛地应用中，在金融领域也不例外。目前，国内至少有 24 只公募基金是基于大数据作为交易策略的基础，规模近 150 亿。除了直接作为交易策略，也可以方便投资者从海量数据中提取有用信息，监控市场舆情。

本文将文本挖掘应用至期货市场中。期货市场天然有应用文本分析的优势，其一，大宗商品在国民经济中的基础地位决定了并不缺乏新闻报道，其二，期货市场活跃的标的数量有限，这易于将文本与标的直接挂钩。

在文本素材准备阶段，目前我们已经抓取了我的钢铁网、广西糖网、镍吧、中国橡胶网和期货日报可以取得的所有的新闻，未来将不断的充实各类品种的新闻来源，力争涵盖全市场。还建立了一个证券期货行业专用的分词词库，基于充实的分词词库才能准确识别文本中的专用词汇。

在市场热点追踪方面，我们认为仅用词频观测是失效的，因为词频依赖于文本量，并且常用词汇的出现频率始终很高，并不利于热点监测，取而代之的方法是用词频排名的上升幅度。这样就可以剔除如“钢铁”、“产量”等常用词汇，抓取到“地条钢”这类炒作主题，除此之外，对于周期性炒作话题“天气”也能起到提前警示的作用。

在缺乏做空机制股票市场中，热度在很大程度上代表着看多情绪，在有做空机制的期货市场中关注度也可以意味着做空，因此用“量”来衡量多空情绪在期货市场中不一定适用，所以需要用文本分析新闻的内在的多空情绪。本文用了构建情感词典的方式逐条分析新闻对标的价格的正负面影响，并计算出新闻情绪指标。基于情绪指标构建了的交易策略，钢铁品种年化收益率可达 23.84%，参数较少，无杠杆调整，还有很大的优化空间。

明显，已有的大数据基金已经在股票市场中取得了亮眼的成绩，主要的依据基础还是以搜索量、关注量、购买量等数量数据，少有涉及文本情感的分析，本文验证了这种分析在期货市场确实可行。不仅在金融领域，在很多前沿领域中让机器理解人类语言的自然语言分析技术都是一项重要课题。

本篇分析是基于情感词库来识别文本中的价格影响，文本的表达方式和词库的完备程度都会对结果产生较大的影响，这是一个局限。另一种基于机器学习的方法，需要大量有情感标注的语料库，如有评级的分析师研究报告。尽管现阶段机器学习对于新闻素材的本文分析适用性不高，但是未来的潜力无限。

期货走势评级体系（以收盘价的变动幅度为判断标准）

走势评级	短期（1-3 个月）	中期（3-6 个月）	长期（6-12 个月）
强烈看涨	上涨 15%以上	上涨 15%以上	上涨 15%以上
看涨	上涨 5-15%	上涨 5-15%	上涨 5-15%
震荡	振幅-5%-+5%	振幅-5%-+5%	振幅-5%-+5%
看跌	下跌 5-15%	下跌 5-15%	下跌 5-15%
强烈看跌	下跌 15%以上	下跌 15%以上	下跌 15%以上

上海东证期货有限公司

上海东证期货有限公司（简称东证期货）是东方证券股份有限公司全资子公司，注册资本达10亿元，系国内四家期货交易所的结算会员。

东证期货专注于金融期货和商品期货的研究与服务，提供权威、及时的研发产品服务和投资策略；专注于信息技术的创新，创建安全、快捷的交易通道，开发多样化、个性化的交易系统；专注于构筑全面的风险管理和客户服务平台。

东证期货管理团队管理经验丰富，业绩出众，在业内享有盛誉。人才管理及激励机制完善，公司拥有硕士学历以上人员占比30%，具有海外证券和期货经历的高端人才占比10%。

2010年，东证期货发展迅猛，成绩斐然，成为业内进步最快、最受瞩目的期货公司之一。2011年初，东证期货荣获2010年度中国金融期货交易所年度会员金奖，同时获投资者教育奖、客户管理奖、技术管理奖和功能发挥奖等四项单项大奖；荣获上海期货交易所优胜会员第七名，铜、橡胶和燃料油三项企业服务奖；荣获大连商品交易所优秀会员第九名；东证衍生品研究院（原东证期货研究所）荣获大连商品交易所、和讯网第二届全国“十大期货研发团队”农产品团队全国第二名、化工团队全国第五名；荣获郑州商品交易所行业进步奖等。

东证期货全年无风险事故，充分体现了公司稳健经营，稳步发展的经营宗旨。

分析师承诺

欧阳静宜

本人具有中国期货业协会授予的期货执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接接收到任何形式的报酬。

免责声明

东证衍生品研究院

地 址： 上海市中山南路 318 号东方国际金融广场 2 号楼 22 楼
联系人：梁爽
电 话： 8621-63325888-1592
传 真： 8621-33315862
网 址： www.orientfutures.com
Email： research@orientfutures.com