# 第五章：基于互联网大数据的资源能源政策经济因素分析

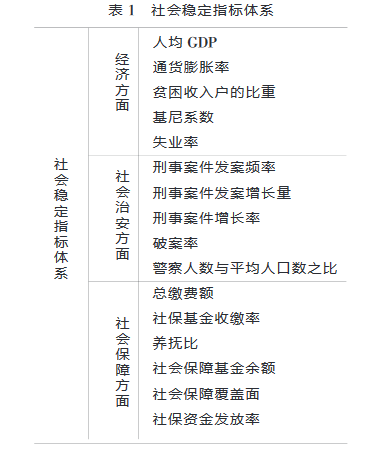
## 5.1国家政治和经济稳定性评估技术分析

### 5.1.1国家政治和经济稳定性评估模型

国家政治和经济稳定是国家社会稳定的重要因素。资源能源生产国的政治和经济稳定性直接关系到资源能源来源的安全和风险，资源能源产品的出口国的政治和经济稳定性也直接关系到我国资源能源的需求和消费。

上海《社会稳定指标体系》课题组（2002）指出，用于衡量社会政治、民生稳定的一般状态指标可以由以下指数定义：社会痛苦指数、公共安全指数、精神卫生指数、社会紧张指数和利益保障指数，其中具体包括通货膨胀率、城镇登记失业率、刑事案件立案率、重大交通事故发生率、每万人精神病患者比率及增长率、信访率、劳动合同纠纷受理件数及增长率等23个指标。同时，还考虑了社会稳定突变状态指标，包括社会冲突、腐败、分裂指数；以及社会稳定解释性指标，包括经济总量、生活水平、贫富差距、社会保障、政治透明度等指数。

夏帆（2004）从三个方面构建社会政治稳定指标体系。包括经济方面的人均GDP、通货膨胀率、失业率、基尼系数等指标；社会治安方面的刑事案件发案频率、增长率、破案率等指标；社会保障方面的社保基金总缴费额、社保基金收缴率、社会保障覆盖面等指标。见下图。



柳长兴，秦琴（2011）从主观、客观两个方面考虑社会稳定水平。客观因素中包括基尼系数、失业率、CPI、刑事案件发案率、公共财政投入、社会保障覆盖率等指标。主观因素则从总体生活、政治生活、经济生活、法治与权益质量等满意度维度指标进行测算，建构了社会稳定的主观综合指标体系，设计科学合理的抽样模型，用问卷调查和社会科学统计分析方法调查计算重庆市社会稳定总体性水平。

综上所述，社会政治稳定水平可以由主观和客观的指标进行衡量，而客观指标更受到研究者的关注。其中受到普遍认可指标的有：经济指标如通货膨胀率、失业率；社会治安指标如刑事案件、重大案件发案频率；社会保障指标如社会保障覆盖面等。

### 5.1.2利用社交媒体大数据进行稳定性评估综述

互联网社交媒体提供给人们随时随地发布、传播信息的服务，其信息传播的广泛性和快捷性使得其成为人们政治、社会生活的新的平台。2010年[阿拉伯世界](http://baike.baidu.com/item/%E9%98%BF%E6%8B%89%E4%BC%AF%E4%B8%96%E7%95%8C" \t "http://baike.baidu.com/_blank)爆发的民主运动“阿拉伯之春”被认为是一场“Twitter革命”。有媒体评论，“阿拉伯之春”属于“谙熟互联网、要求和世界其它大部分地区一样享有基本民主权利的年轻一代”，互联网、社交媒体的重大作用意义可见一斑。

Howard et al 指出，在阿拉伯之春事件中，社交媒体上一定类型的活动加速了大规模抗议等事件的发生。

Sullivan.2011分析了突尼斯同摩尔多瓦、伊朗的抗议活动都可被称为“Twitter革命”的原因。抗议活动期间，传统媒体完全被政府控制失去了传媒功能，无法从中获取事态的最近进展，因此人们转向了YouTube上的每日活动视频, 以及 Twitter、Facebook等社交媒体上的博客信息。特别是由于网络站点封锁和网络攻击等原因，政府难以对互联网上的抗议信息进行控制，导致相关信息广泛传播，网络上抗议呼声一时高涨。

Bruns, Axel,Highfield, Tim, &Burgess, Jean E.(2013）也考察了在埃及和利比亚的冲突中，英语和阿拉伯语用户在Twitter上的发言情况。文章通过统计微博数量分析得出结论，阿拉伯语用户中，埃及的Twitter活跃用户数相对较高，总微博数量也达到一定规模，因此可以被用于观察事件参与者的情感倾向，同时也证实了，以Twitter为代表的网络社交媒体在埃及的国家抗议活动中起到了一定传播、组织和发布信息的作用。

上述研究均分析了“阿拉伯之春”在社交媒体上的强烈反响，探讨了社交媒体在事件中起到的作用。正因如此，社交媒体上的动态也可以作为分析、预测实际事件走向的依据。

Xiaofeng Wang et al. 2012使用Twitter微博文本数据对肇事逃逸案件的发生进行预测，将犯罪预测的监督学习框架和社交媒体数据进行结合。文章整体采用有监督学习框架，使用一定量的微博文本和对应时间段发生的肇事逃逸案件信息进行学习，需要采集微博数据，和真实的案件发生数据。其中使用了较为精确的自然语言处理工具来对微博文本进行特征提取，首先NLP（natural language processing）工具抽取一条微博中的事件内容，再使用半监督的基于事件的主题分类模型确定关键的主题词，并根据主题词来进行预测。单纯采用时间维度的分析并不能有效提取案件的发生规律，预测效果近似随机猜测，而该文中的方法则能显著提升预测性能。

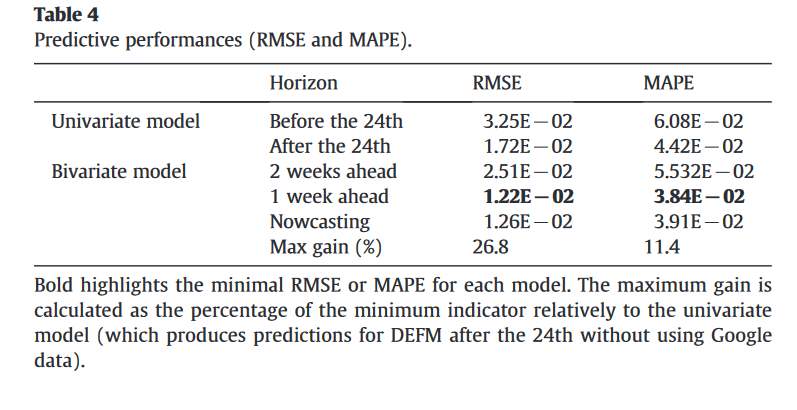
目前研究仅针对了肇事逃逸案件，如能获得实际的案件信息，也可用于其他类型的案件，特别是对国家政治稳定性有重大影响的案件，如抗议示威、大规模游行等案件，对于评估社会稳定突变状态指标，或刑事案件发案率等社会治安指标都有一定作用。

### 5.1.3利用搜索引擎大数据的评估综述

由于越来越多的经济活动都转变为在网络上进行，或者在网络所提供的种种服务中留下交易痕迹。因此搜索引擎在评价经济指标上的作用得到了研究者的重视，一系列研究都表明以Google Trends为代表的搜索引擎大数据对阐释经济市场的波动震荡，和做出精确地经济评估及预测有明显的作用。

Askitas and Zimmermann (2009) 发现了Google Trends搜索量和德国的失业率之间有强正相关性。Choi and Varian(2009b) 对美国，Suhoy (2009) 对以色列，D'Amuri (2009)对意大利 ，以及D'Amuri and Marcucci (2009) 对美国的失业率进行考察时都有相似的结论。

Fondeur and Karamé (2013) 验证了 Google Trends数据能否用于预测法国青少年（15-25岁）的失业人口数量。文章使用了政府发布的周度和月度失业人口数据，结果显示，使用了 Google Trends数据的预测模型效果要远远优于未使用的模型。在两年内的失业数据上验证，使用了 Google Trends数据的预测模型对预测质量整体提高了27%，对预测的准确度平均提高了40%。（见下图）



利用搜索引擎大数据的评估主要从两个思路来提升性能，一是搜索关键词的选择，另一是建模方法的使用。在关键词选择中，Fondeur and Karamé (2013)简单选择了EMPLOI（意为工作、失业），文章认为，该关键词可以作为劳动力市场供需双方查找信息的入口。Askitas and Zimmermann (2009)选择了四组，每组1-8个关键词，如HR咨询、招聘网站，以及直接使用失业、就业等词语，从4个角度对失业相关的搜索入口进行了考察。Choi and Varian(2009b) 选择了jobs、welfare & unemployment两个关键词。D'Amuri and Marcucci (2009) 仅选择了jobs一个关键词。可见，直接使用失业、就业等词语，选择较为简单而少数的关键词是目前采用较多的方法。在建模方法上，Fondeur and Karamé (2013)选择了扩散的Kalman滤波和最大似然估计法，以得到非参数滤波不能达到的更好的数据贴合度。

## 5.2重要公投与政策预测技术分析

### 5.2.1研究现状综述

英国脱欧、苏格兰独立等国际上的一些重要的公投事件对于国际政治、经济、金融和外交环境都有着密切的联系，对资源能源产品的生产和出口也有直接或间接的影响。公投结果的预测方法受到了广泛关注。

Ronald MacDonald 等提出了一种使用Google搜索引擎数据和计量经济学思路分析苏格兰独立公投的方案，使用Google Trends中与公投相关的周搜索量数据建立计量经济学VAR模型，在2013年8月至2014年9月公投前2天的数据上预测赞成票率为45%，与实际结果44.7%十分接近。利用Google Trends日搜索量数据对公投前最后一个月的影响预测结果的因素进行分析，定量地探讨了期间重大事件和其他因素对结果的影响程度，为合理解释公投过程和民众心理提供了新的视角和依据。

传统的公投预测使用的新闻报道等信息是对提供给投票者的信息进行分析。 这种信息是投票者被动接受，并不能从中了解投票者的情感倾向。而利用Google Trends 数据则可以分析投票者主动希望了解的信息。

### 5.2.2利用搜索引擎大数据的预测技术路线

利用Google等社交媒体数据进行选举预测的技术路线可以总结如下：数据采集、预测方法应用和效果评价。下面针对前两个部分进行详细阐释。

1. **数据采集**

利用Google Trends等搜索引擎工具提供的搜索量查询接口，设定一定的关键字和采集时间范围以得到搜索量数据，还可以设置地理位置的限制。

Ronald MacDonald 等在利用Google Trends对苏格兰公投进行预测时，采用单一的关键字“Alex Salmond”，采集时间为2013年8月至2014年9月公投前2天，同时设置了搜索请求的地理位置必须在苏格兰境内。

其中搜索关键字的选择很大程度上决定了预测的质量，在Ronald MacDonald等在苏格兰公投的案例中，提出了选择搜索关键字的两个原因，为关键字的选择提供了一定思路：

1. Alex Salmond为公投时期苏格兰首席大臣，并且担任苏格兰民族党党首（主张苏格兰独立），因此是投票者最可能在Google 上搜索以了解苏格兰独立相关信息的关键词。
2. 在苏格兰议会选举中，“Alex Salmond”的搜索量就和最终的投票表决结果有高度的正相关性。如其在2011年2月到5月间的搜索量和投票结果有0.6的相关性。
3. **预测方法应用**

Ronald MacDonald 等使用Google Trends中该关键词的周搜索量数据建立计量经济学VAR模型，并在考虑了意外事件波动时对结果进行了适当修正。VAR模型及建立过程采用了Brooks and Tsolacos(1999) 与Vozlyublennaia (2014)中实用的方法。采用的两个主要变量分别是“可能的赞同票数”（来自6家主要选举公司的数据平均）和Google Trends搜索量。修正结果时考虑了变量的不确定性进行动态预测。Doornik and Hendry (2013)也采用了类似的模型和流程流程。

## 5.3重要选举预测技术分析

### 5.3.1研究现状综述

资源能源生产国、资源能源产品的消费国、资源能源运输重要通道的所在国，根据其政治体制的不同，大都会周期性的产生其国家或地区的政治领导人。新领导人的政治倾向和政策选择将会对资源能源的安全风险产生巨大的影响。重要选举的预测也逐渐从仅仅依靠传统的民意调查到同时利用互联网大数据提供的社会舆论信息。

在选举预测问题中Twitter数据得到了广泛应用。O’Connor et al.2010最早使用Twitter数据代替传统的民意调查，根据预先得到的词汇与其情感的对照表给每条微博的情感评分，再对正面、负面的微博数量统计。将微博中针对某个候选人的情感倾向评分与选票结果进行关联度分析，探讨了微博数据的选举预测作用。Tumasjan et al.2010对Twitter数据中提及选举各方政党的微博数量进行统计，将结果用于预测2009年德国联邦选举，预测的胜率误差为1.65%，效果优于传统民调。Bermingham and Smeaton 2011利用Twitter数据，并使用了多种语义分析方法，在2011年爱尔兰大选中进行测试，预测的胜率误差为3.67%，略优于民调结果。Metaxas et al.2011使用Twitter中参选者被提及数量以及语义分析的方法对2010美国各州选举进行预测，在6个州中两种方法均有3个州正确预测了获胜方。

也有研究者探讨了搜索引擎数据在选举预测中的使用。C. Lui 2011关注了 Google搜索量数据对于2008和2010美国国会选举的预测作用。他们根据候选人之间竞争的激烈程度将全部选举划分为若干竞选单元，发现使用Google搜索量数据的预测在部分单元中得到了比随机猜测更好的效果，最好结果出现在2008年的某个单元，达到了81%的准确度。[17] C. Lui, P. T. Metaxas, and E. Mustafaraj, “On the predictability of theu.s. elections through search volume activity,”

### 5.3.2利用社交媒体大数据的预测技术路线

作为目前研究选举预测问题的最常见思路，利用Twitter等社交媒体数据进行选举预测的技术路线可以总结如下：数据采集、数据清洗、预测方法应用和效果评价。下面针对前三个部分进行详细阐释。

**1）数据采集**

主要利用Twitter等社交媒体提供的数据访问接口，如Twitter search API来获取选举前一段时间内全部或随机比例的微博用户发布的与选举相关的微博信息，从选民个人发布的相关内容中推测个人及选民整体的选举意向。其中需要确定采集的时间，过滤微博的关键字及其他参数配置。

现有相关研究在数据的采集时间上有较多差别，虽然采集的截止日期大多选择了选举日前一天，但是开始时间却各不相同，时间跨度从选举前一周到选举前三年均有实验。目前时间选择对预测结果的影响尚没有确切的结论。如Tumasjan et al.2010在预测2009年德国联邦选举时，选取了选举前7周到选举前1周的数据。而Tjong Kim Sang& Bos 2012在预测2011年荷兰参议院选举时则只选取了选举前1周到选举日前一天的数据。

同时，大部分研究都对微博内容进行了一定过滤，限制内容的题材为与待预测的选举相关,通常会选择候选人的姓名、参选政党的名称作为过滤的关键词。Metaxas et al.2011在预测2010年美国选举时采用了候选者姓名在gardenhose数据库中过滤相关微博条目。Skoric et al.2012在预测2011年新加坡大选时同时采用了候选人的姓名和参选政党的名称过滤相关微博条目。

1. **数据清洗**

由于选民的微博是目前多数研究采用的数据来源，因此以下讨论的数据清洗也主要针对这种情况。

* 保证数据纯粹性

保证数据的纯粹性要求发布微博等社交信息的用户是可能的参选者。鉴于这种信息并没有在用户个人资料中明确显示，因此需要通过其他信息来推断。目前研究中所采用的限制措施有两种，即限制地理位置和限制微博内容的发布语言。

限制地理位置时可以通过控制Twitter search API的有关参数，Gayo-Avello 2011在预测2008年美国大选时就采用了设置参数的方法将微博数据来源限制在所希望了解的郡县中。也可以通过限制微博发布者的地理位置达到上述目的，Skoric et al.2012在预测2011年新加坡大选时限制了发布者的位置必须在新加坡境内以去除不具有选举资格的用户发表的内容。

限制发布语言主要应用在德语、荷兰语等全球化程度不太高的语种上，如Tumasjan et al.2010 以及Jungherr et al.2011在预测2009年德国联邦选举中都限制了微博必须使用德语，以减少数据集中非德国民众发布的微博信息，在一定程度上达到了与限制地理位置同样的效果。

* 去除数据偏差

去除偏差即减少数据的不平衡性，使之可以成为统计意义上的代表样本。Twitter等社交媒体中采集到的用户群并不一定符合人口统计学上的均衡性，而这种偏差导致了数据样本的片面性，特别是在选举这种需要考虑全局民意的情况下会造成对预测结果较为严重的的人为干扰。解决这种数据偏差的一个直观思路是对每个用户所属的人口统计学上的社会阶层进行划分和判断，根据用户所在阶层的选举活动参与程度得出其发布内容的重要性，给予不同用户的信息不同的的权重。

然而目前的互联网社交媒体中没有强制用户登记他们完善的真实信息（如性别、年龄等），而从用户发布的内容中推测这些个人资料又涉及大量工作，因此使得将用户划分为不同的投票阶层十分困难。但仍然有研究者做出了一定尝试，Gayo-Avello 2011 利用用户登记的真实姓名、郡县等信息，以及线上的公开查询记录得到了2500名用户的年龄，以此估计出他数据集中用户的年龄范围在18-44周岁之间。接下来，他根据在2004年大选中选民的年龄分布情况对数据集中的用户信息给予不同的权重。这两步措施使得预测误差从13.10%降低到了11.61%，可见进行了数据去偏差后对性能带来了显著提升。

* 数据去噪声

数据去噪声表示去除与实际选民的选举意向无关的内容，如没有信息量或灌水、炒作的条目。数据中噪声的存在会降低预测方法的使用效率，更有可能对预测过程造成干扰，使预测结果出现错误。尽管目前的选举预测研究中没有研究者采用过实际的数据去噪声方法，但一些研究者已经意识到了数据去噪的重要性和必要性。

Metaxas et al.在2011年的研究中指出，恶意灌水和政治宣传者都会利用软件创造一些虚假的社交媒体账户，并通过这些假账号夸大、散布他们的政治主张，刻意营造其政治势力，这种行为污染了每个数据集观察者的视听。因为社交媒体所呈现的信息在人群中有广泛的影响力，这一平台对恶意灌水和政治宣传者有重要意义，使他们很可能在其中混淆视听，达到干扰舆论的效果。

更进一步，他们进行了实验以验证基于词极性的语义预测方法在被灌水信息上的鲁棒性表现，结果显示，仅仅依靠词极性的语义预测方法并不能识别出被灌水信息和正常微博信息的微妙区别，并且会对这些虚假的信息作出错误的判断。更加表明了，在选举预测过程中进行数据去噪声的步骤应该是必须完成的。

1. **预测方法应用**

使用Twitter进行选举预测方法的核心即是从微博中推测选民意向，可以先推测单个用户的意向再对整体进行统计，也可以跳过单个用户，直接以整体意向为单位统计、推测。预测方法在范围上可分为在国家、州、郡县等层面，在预测结果上可分为预测获胜率或预测获胜方。

从技术角度，预测的方法主要有以下两类：

* 使用关键词提及数量

简单统计微博中提及某一选举人或政党的数量，预测的准确率随着总微博数量的增大而提高。其优点在于：易于实现；可以在实际临近投票结果前应用；不仅可以用于预测选举的总体情况，并且基于用户倾向选举的对象会在其微博中被最多提及还可以用于预测单个用户的选举倾向；并且被认为是与传统民意调查的思路接近而具有很高的准确率。

Tumasjan et al.2010 以及Jungherr et al.2011在预测2009年德国联邦选举中都采用了微博关键词计数的方法。Tumasjan et al.2010 对提到参选各党派的微博进行统计，发现各党派微博数量的排名与其在选举中得票数的排名相同，并且有相似的比例。简单地采用微博数量进行预测可以达到误差1.65%的效果，与6家专业选举机构发布的预测结果比较，表明预测水平近似。这一结果与Véronis (2007)指出观点相符合的，即在2007年法国总统选举中，简单地使用参选者提及数统计可以获得比传统民调更好的结果。

Tjong Kim Sang& Bos 2012在预测2011年荷兰参议院选举时遵循了与Tumasjan et al.2010类似的思路，他们根据微博用语的特点，同时对政党名称的缩写及全称进行计数，按照政党名称的提及数量占比预测其获得的席位，与2家选举公司的结果对比，达到了相似的结果。进一步，他们分析了数据集中存在的偏差和噪声问题，只保留了每个用户的第一条、且只提及一个政党的微博，获得了更好的效果。

但由于其未考虑任何语义信息，忽视了文本中不同的感情倾向而受到质疑。Gayo-Avello 2011在预测2008年美国大选时的实验显示，单纯使用微博数量统计的预测效果甚至远远低于随机猜测。

* 使用语义情感分析

只有少数研究采用机器学习的方法对文本进行正面、负面的情感分类如Bermingham and Smeaton 2011 ，Tjong Kim Sang and Bos 2012。

多数研究仅使用了词汇特征，根据词汇与其所对应感情（词汇极性）的对照来推测整条微博的情感和选举倾向。这种方法简单，易于实施，但也被证实有许多分类错误，精确的语义情感分析仍然是一个困难的挑战。

O’Connor et al.2010在预测2008年美国大选时，首先筛选出包含与选举相关关键词的微博条目，如包含obama和mccain的条目。再根据词汇极性对照表OpinionFinder（Wilson et al.2005），只要一条微博信息中出现了正面或负面情感的词汇，就相应地认为该条微博针对该关键词也具有正面、负面的倾向。 为了衡量某天的正面情感的强度，定义分值为当天内该关键词正面微博条目数/负面微博条目数。

Gayo-Avello 2011在预测2008年美国大选时，采用了两种基于词极性的方法。一种方法对微博中词极性进行统计，如果正面词汇多于负面词汇，则认为该条微博也有正面倾向。通过统计某个用户对于所有候选者的正面倾向评分，可以得出分数最高者即为该用户的选举意向。一种为Vote & Flip方法，首先对句子中不同极性的词汇进行统计，再应用一系列复杂规则来判断整个句子的极性，对传统的简单情感推测进行了规则上的扩充。

该研究还采用了另一种semantic orientation（语义源分析）方法。该方法在句子中寻找有特定情感倾向的短语，这些短语预先定义的短语可以借助搜索引擎获得，也可以采用那些已经表现出自己选举意向的微博中短语。其基本思路与词极性法相似，而将最小的情感推测单元扩充到了短语层面。