# 短文本中的情感强度检测[[1]](#footnote-2)

迈克·瑟沃尔，凯旺·巴克利，圣乔治·帕尔托格鲁，迪·蔡

英国伍尔弗汉普顿WV1 1SB伍尔弗汉普顿大学计算与信息技术学院统计网络计量学研究组。

邮箱:m.thelwall@wlv.ac.uk，K.A.Buckley@wlv.ac.uk，G.Paltoglou@wlv.ac.uk，caid@wlv.ac.uk

电话:+44 1902 321470传真:+44 1902 321478

阿尔维德·卡帕斯

不来梅雅各布大学人文社会科学学院，校园一环，

德国不来梅28759号

电邮:a.kappas@jacobs-university.de

电话:+49 421 200-3441

**每天都有大量的非正式信息发布在社交网站、博客和论坛上。在这些文本中，情感对于表达友谊、显示社会支持或作为网上争论的一部分似乎经常很重要。需要识别情感和情感强度的算法来帮助理解情感在这种非正式交流中的作用，并且还需要识别不适当或异常的情感话语，这些情感话语可能与对自己或他人的威胁行为相关联。然而，现有的情感检测算法往往是面向商业的，旨在识别对产品的意见，而不是用户行为。本文通过一种新的算法SentiStrength部分填补了这一空白，该算法使用新的方法来利用网络空间事实上的语法和拼写风格，从非正式英语文本中提取情感强度。SentiStrength应用于MySpace评论，并使用机器学习优化的术语情绪强度查找表，能够以60.6%的准确率预测积极情绪，以72.8%的准确率预测消极情绪，两者都基于1-5的强度等级。前者优于基线和各种通用机器学习方法，但后者不行。**

# 介绍

大多数观点挖掘算法试图识别文本中情感的极性:积极、消极或中性。虽然对于许多应用来说这已经足够了，但是文本通常包含正面和负面情绪的混合，并且对于一些应用来说，有必要同时检测这两种情绪，并且还检测所表达的情绪的强度。例如，监控在线交流中的情绪的程序，可能被设计为在使用不适当的情绪时进行识别和干预，或者识别处于风险中的用户(例如，Huang，Goh，& Liew，2007)，将需要对所表达的情绪的强度以及参与者是否适当地平衡积极和消极的情绪保持敏感。此外，了解情绪在在线交流中的作用的基础研究(例如，Derks、Fischer和Bos，2008；例如，Hancock，Gee，Ciaccio和Lin，2008年；纳尔迪，2005)也将受益于细粒度的情绪检测，正如越来越多的心理学和其他社会科学研究情绪在各种类型的讨论或一般话语中的作用(Balahur，Kozareva和Montoyo，2009；彭尼贝克、梅尔和尼德霍夫，2003年；肖特和帕尔默，2008年)。

在线情感检测的一个复杂因素是，在许多电子通信媒体中，基于文本的英语通信似乎经常忽略语法和拼写规则。也许最著名的是手机文本语言及其缩写，表情符号和截断的句子(格林特&埃尔德里奇，2003；Thurlow，2003)但类似的风格在许多其他形式的计算机媒介交流中也很明显，包括聊天室、公告栏和社交网站(Baron，2003；Crystal，2006)。被广泛认可的创新包括表情符号，如:-)，它在传达情感方面相当有效(Derks，Bos和von Grumbkow，2008；Fullwood & Martino，2007)和单词缩写如m8 (mate)和u (you) (Thurlow，2003)。虽然有时被视为糟糕的语言使用，但这是对与系统相关的技术能力和社会因素的自然反应(Baron，2003；瓦尔特&帕克斯，2002)。这些变化引起问题，因为典型的语言情感分析程序从词性标注(例如，Brill，1992)开始，词性标注依赖于标准拼写和语法，和/或应用假设即使没有正确的语法，至少也是正确的拼写的规则。拼写纠正在这种情况下可能是有用的，但这是基于拼写偏差很可能是偶然错误的假设(Kukich，1992；Pollock & Zamora，1984)，因此当前的算法不太可能很好地处理故意非标准的拼写。然而，有一系列常见的缩写和新词是语言学算法原则上可以检测到的。非语言机器学习算法通常基于文档中单个单词、单词对和单词三元组的出现来预测情感。由于拼写问题和情感表达的创造性，这些方法在非正式文本上的表现也可能很差，即使有大的训练语料库可用(见下文)。

社交网站MySpace，当前研究中使用的数据来源，因其年轻成员、音乐取向和非正式交流模式而闻名(boyd，2008；博伊德，2008年)。可能由于这些因素，朋友之间交流的95%的英语公开评论包含至少一个来自标准英语的缩写(Thelwall，2009)。常见的特征包括表情符号、短信式的缩写以及使用重复的字母或标点符号来强调(例如，loooong time，Hi！！！).评论通常很短(平均18.7个单词，中位数13个单词，68个字符)(Thelwall，2009)，但积极的情绪是常见的(Thelwall，Wilkinson和Uppal，2010)。

本文提出了一种新的算法SentiStrength，该算法采用多种新颖的方法从简短的非正式电子文本中同时提取正面和负面情感强度。SentiStrength使用带有相关强度测量的情感词词典，并利用一系列公认的非标准拼写和其他表达情感的常见文本方法。SentiStrength是通过最初的2600条人类分类的MySpace评论集开发的，并在1041条MySpace评论的随机样本上进行评估。请注意，在一些文章中，而不是在情绪心理学中，情绪一词指的是分为积极、消极和中性的情感，而情绪一词指的是更有区别的情感(例如，快乐、悲伤、恐惧)。相反，这两个术语在这里作为同义词使用，它们的含义由下面描述的编码器指令有效地定义。本文的主要创新贡献是:优化情感术语权重的机器学习方法；非正式文本中重复字母非标准拼写的情感提取方法:以及相关的拼写纠正方法。此外，该论文引入了用于正面和负面情感的双5点系统、用于该系统的1，041个MySpace评论的语料库、以及结合了新颖和现有方法的新的整体情感强度检测系统。

# 背景和相关工作

本文献回顾部分讨论了相关的观点挖掘/情感分析研究以及情绪心理学的一些相关贡献。

## 观点挖掘

意见挖掘，也称为情感分析，是从(非结构化)文本中提取积极或消极的意见(Pang & Lee，2008)。意见挖掘的许多应用包括从多个在线评论中检测电影受欢迎程度，以及通过车主在专门网站或论坛上的评论来诊断他们喜欢或不喜欢车辆的哪些部分。也有与营销无关的应用，如区分情感和信息社交媒体内容(Denecke & Nejdl，2009)。

意见挖掘通常分为两个或三个阶段，尽管某些任务可能需要更多的阶段(例如Balahur等人，2010)。首先，输入文本被分成几个部分，例如句子，每个部分被测试以查看它是否包含任何情感:它是主观的还是客观的(Pang & Lee，2004)。其次，分析主观句以检测它们的情感极性。最后，可以提取表达意见的对象(例如，Gamon，Aue，Corston-Oliver和Ringger，2005)。意见挖掘通常只处理正面和负面情绪，而不是离散的情绪(例如，快乐、惊喜)，不检测情绪强度(但有时使用单词与正面或负面情绪的关联强度，例如，Kaji & Kitsuregawa，2007)，并且不同时识别正面和负面情绪。然而，这种意见挖掘研究可以帮助同时评估积极和消极的情绪强度，因为它对情绪分析的一般见解，也因为大多数技术在理论上可以重新用于这项新任务。例如，短语分析技术可以用于识别积极和消极的情绪，甚至在单个句子中(Choi & Cardie，2008；威尔逊，2008；威尔逊、维贝和霍夫曼，2009年)。

意见挖掘算法通常使用机器学习来识别与积极和消极情绪相关联的一般特征，其中这些特征可以是文档中的单词、词性或n元语法的子集(即，所有n个连续单词的出现频率，其中n通常是1、2或3) (Abbasi，Chen，Thoms，& Fu，2008；Ng、达斯古普塔和阿里芬，2006年；唐，谭，程，2009)。其他成功使用的功能包括:在线电影评论中的表情符号(Read，2005)，它似乎比文字更独立于领域；词汇句法模式(如Riloff & Wiebe，2003)；以及源自形容词极性列表的人工特征(Ng等人，2006)。这些额外的功能通常会在性能上带来微小但显著的提升。基于规则的方法也被用于识别句子中与情感相关的结构(Prabowo & Thelwall，2009；吴，庄，&林，2006)。两个重复出现的机器学习问题是特征选择和分类算法选择。

特征选择，即去除最不有用的n-gram的数据处理，已被证明可以略微提高分类性能，例如，通过选择在信息增益(Riloff，Patwardhan和Wiebe，2006年)或对数似然性(Gamon，2004年)等指标上得分最高的一组受限特征(例如，5000)。当使用n-grams(和词汇句法模式)时，也可以通过修剪由具有更强信息增益值的更简单特征所包含的特征集来进行小的改进(Riloff等人，2006)。例如，如果“爱”具有比“我爱”高得多的信息增益值，则二元模型可以被消除，而没有随后分类的功率损失的太大风险。熵权遗传算法也可以比标准的特征约简方法执行得更好(Abbasi，Chen和Salem，2008)。

在分类算法方面，支持向量机(SVM)被广泛使用(Abbasi et al .，2008；阿巴西等人，2008年；Argamon等人，2007年；加蒙，2004；米什内，2005年；Wilson，Wiebe和Hwa，2006年)，因为它们在大多数机器学习环境中的表现似乎与其他方法一样好，甚至更好。然而，除了少数例外(Read，2005；Wilson et al .，2006)，与其他方法的明确比较还没有包括在意见挖掘出版物中。

许多其他方法也被用于检测文本中的情感。一个是有一个积极和消极词汇(例如，爱，恨)的字典，如在普通问询者(斯通，邓菲，史密斯和奥吉尔维，1966)，WordNet情感(斯特拉帕拉瓦和瓦利图蒂，2004)，SentiWordNet (Baccianella，Esuli和Sebastiani，2010；Esuli & Sebastiani，2006)或Q-WordNet(ager ri & garcía-Serrano，2010)，并计算它们出现的频率。这种方法的修改包括识别否定术语(Das & Chen，2001)，增强情绪的词(例如，真的爱，绝对恨)和整体句子结构(Turney，2002)。一种更复杂的方法是识别在某些上下文中可能是主观的文本特征，然后使用上下文信息来决定它们在每个新的上下文中是否是主观的(Wiebe，Wilson，Bruce，Bell和Martin，2004)。

另一种观点挖掘技术主要使用语言学方法:基于组合语义(基于周围文本的关于单词可能含义的信息)的简单规则来检测表达的极性(Choi & Cardie，2008)。这对于新闻专线文档中被手动编码为至少具有中等积极或消极情绪的短语给出了良好的结果。这种方法似乎特别适合有大量语法正确的文本可以从中学习规则的情况。然而，一项对在线客户反馈中语法质量差的文本的研究表明，当添加到单词袋(1-grams)方法时，语言学方法可以稍微改善分类，尽管积极的特征减少与添加语言学特征具有类似的影响(Gamon，2004)。这种改善可能是因为有大量的数据集可用(40，884个文档，每个文档平均2.26个句子)，正如以前对非正式文本的分析所声称的那样(Mishne，2005)。另一种方法使用评价形容词(例如，“某种程度上”、“非常”)的词典以及取向词典来检测电影评论的极性。其性能不如unigrams，但综合性能优于单独的unigrams(arg amon et al .，2007)。语言特征也被成功地用于将观点挖掘扩展到一个多方面的变体，该变体能够检测关于主题的不同方面的观点(Snyder & Barzilay，2007)。一个有希望的未来方法是结合关于使用情感的原因的背景，例如区分意图、争论和推测(Wilson，2008)。

## 检测多种情绪

情绪心理学研究认为，虽然积极和消极情绪是重要的方面，但有许多不同的广泛被社会认可的情绪类型，情绪的强度(唤醒水平)可以变化(例如，Cornelius，1996；福克斯，2008)。在来自心理学的情绪维度模型(Russell，1979)中，情绪总是可以从根本上分裂为两个轴:唤起(由低到高)和效价(由积极到消极)。虽然这个模型是有用的，但其他研究表明，积极和消极情绪可以共存(例如，Fox，2008，第127页),并且在许多情况下是相对独立的——特别是当情绪水平不极端并且持续时间较长时(Diener & Emmons，1984；胡珀特和惠廷顿，2003年；沃森，1988；Watson，Clark和Tellegen，1988)，因此，将情绪视为可单独测量的积极和消极成分似乎也是合理的，正如一个流行的心理学研究工具所编码的那样(Watson等人，1988)。

之前已经有一些尝试来开发算法，以检测文本中情绪或情感的强度或流行度，或者区分几种类型的情感。例如，心理学的LIWC(语言学调查和字数统计，www.liwc.net)软件使用一系列带有情绪的单词来检测文本中的积极和消极情绪，以及在心理学和心理治疗中特别有用的三种特定情绪:愤怒、焦虑和悲伤。它使用简单的单词计数，测量落在广泛的预定义列表中的单词的比例(例如，408个正单词和499个负单词或词干)。该列表包括一些与情绪相关但不描述情绪的单词。例如,“幸运”是一个肯定的关键字，而“失败”是一个否定的关键字。与上面讨论的机器学习方法相反，这些列表是使用人类评委小组和统计测试来编译和验证的。

LIWC计算文本中情感的流行程度，而不是试图诊断文本的整体情感或情感强度。它最适合于较长的文档，因为它的统计数据是情感发生趋势的有用指标。为了简单起见，该程序使用了单词截断(例如，joy\*匹配任何以joy开头的单词)，而不是词干化或词汇化，但没有考虑像“very”这样的助词或否定的否定效果(例如，不快乐)。心理学研究者使用LIWC来研究语言和心理之间的联系(Pennebaker等人，2003年)，也将其作为一种实用工具，例如，根据人们的语言使用来检测他们应对丧亲之痛的能力(Pennebaker，Mayne和Francis，1997年)。一种相关的情绪检测方法基于学生描述他们日常生活所使用的词语来区分快乐、不快乐和中性状态(Wu等人，2006)。然而，这类似于意见挖掘的典型的正面/负面/中性目标。

一项计算机科学倡议试图识别文本中的各种情绪，集中于六种所谓的基本情绪(Ekman，1992；Fox，2008)的愤怒、厌恶、恐惧、喜悦、悲伤和惊讶(Strapparava & Mihalcea，2008)。这项倡议还测量了情感强度。使用人类注释的语料库，编码者为每个文本(新闻标题)的每个情感分配从0到100的强度，尽管注释者之间的一致性很低(皮尔逊相关度为0.36到0.68，取决于情感)。随后在这个数据集上训练了各种算法。例如，有人使用WordNet情感列表为六种情感生成合适的词典。第二种方法使用了朴素贝叶斯分类器，该分类器在LiveJournal博客集上进行训练，这些博客由它们的所有者用六种情绪之一进行注释。最好的系统(用于细粒度评估)是以前为报纸标题设计的系统，UPAR7 (Chaumartin，2007)，它使用语言解析和标记以及WordNet、SentiWordNet和WordNet Affect，因此依赖于合理正确的标准语法和拼写。

在心理学中，情绪一词指的是中期和长期的情感状态。一些博客和社交网站允许成员在编辑状态或写文章时描述他们的心情，通常是通过从一系列图标中进行选择。这些结果可以用作带注释的语气语料库。理论上，这样的语料库应该可用于训练分类器，以从与情绪图标相关联的文本中识别情绪，并且已经设计了一个系统来做到这一点，但是成功有限，可能是因为所分析的文本通常很短(平均200个单词)并且有许多情绪，其中一些情绪彼此非常相似，尽管即使是二元分类任务也具有有限的成功(Mishne，2005)。一个后续项目试图使用从所有文本的集合中得出的199个单词(1-grams)和单词对(2-grams ),而不是通过对单个文本进行分类，来得出在特定时间段内具有给定情绪的帖子的比例(Mishne & de Rijke，2006)。结果显示与总体自我报告情绪高度相关。类似的汇总方法随后被应用于一系列社会科学领域(Hopkins & King，2010)。

语言处理还与预先存在的大量主观常识陈述模式相结合，并应用于电子邮件消息中相对非正式和领域无关的文本，以检测多种情绪(刘，Lieberman，& Selker，2003)。然而，这是电子邮件支持系统的一部分，情绪检测的准确性没有被直接评估。

## 情感强度检测

除了上面讨论的关于多种情绪强度检测的研究(Strapparava & Mihalcea，2008)，还有一些关于积极-消极情绪强度检测的工作。先前的一项研究使用改进的情感分析技术来预测人类对电影评论的评分强度，评分等级为1到5(Pang & Lee，2005)。这是一种情感强度评估，具有正面和负面情感的组合尺度。对人类判断的实验导致作者合并了两个类别，因此最终的任务是4个类别的分类，同时为了测试目的还构建了3个类别的版本。多类SVM分类法与SVM回归法的比较表明，当全部4个类别都被使用时，SVM回归法比多类SVM分类法略好，但当仅使用3个类别时，回归法就不行了。随着类别数量的增加，SVM回归的相对性能似乎会进一步提高，因为类别的排序是隐含信息，多类别SVM不使用，但SVM回归使用。当每篇评论中增加了关于肯定性句子的百分比的信息时，也有轻微的改善。然而，这可能与非常短的文本的语料库无关。

情感强度分类也已经被开发用于新闻专线文本中的主观句子或从句的三级方案(低、中、高或极端)，其使用将句子转换成反映其结构的依存关系树的语言分析(Wilson等人，2006)。与单个词元相比，将依存关系树添加到词元大大提高了各种分类器的性能，这可能是由于相当大的训练集(9，313个句子)、(大概)高质量的文本语法以及在这项任务上相当低的初始性能(对于词元为34.5%到50.9%，对于应用于1级子句的三种类型的分类器，上升到48.3%到55.0%)。在这里，SVM回归被基于规则的学习Ripper (Cohen，1995)和BoosTexter(一种结合了多个弱分类器的增强算法)所超越。

与目前的论文非常相似的是一个使用多种方法的组合测量与对话系统相关的非正式文本中的多种情绪及其强度的论文，包括通过重复的标点符号(例如！！)，表情符号和大写字母以及翻译缩写(Neviarouskaya，Prendinger，& Ishizuka，2007)。该系统还在0-1的范围内测量情绪强度，并使用由三名人类法官指定的术语和强度评级字典(符合率中等:Fleiss Kappa 0.58)。据报道，对160个人类编码的句子的评估表明，在68%的句子中，系统与编码者的平均误差在20%以内。

# 数据集与人类对情感强度的判断

MySpace被选为本次研究的测试数据来源，因为它是一个包含大量非正式文本语言的公共环境，并且作为2009年世界上访问量最大的网站之一，它本身就很重要。MySpace评论的随机样本是通过检查2007年6月18日加入的第15个成员的个人资料进行的，多达4万人，并选择那些声明拥有美国国籍且公开资料不是音乐家、喜剧演员或电影制片人的人。其中，朋友少于两个或没有评论的人被视为不活跃，而朋友超过1000人或4000条评论的人被视为不正常。然后为每个剩余成员确定一个评论朋友，满足上述相同的标准，并从两者之间的每个交流方向随机选择一个评论。这些意见摘录于2008年12月。这产生了一个美国评论者-评论者信息的大型基本随机样本。垃圾评论和连锁消息随后被清除，包含图像的评论也是如此。

尽管情感分析通常涉及观点(Pang & Lee，2008)，但Wilson (2008)将其概括为从文本中识别作者隐藏的内部状态的心理任务。然而，对于MySpace的数据来说，目的不是确定观点或作者的内心状态，而是确定在线交流中表达的情感的作用。因此，任务的重点是识别每个信息中表达的情绪，无论是反映作者隐藏的内部状态，预期的信息解释，还是读者隐藏的内部状态。

为了获得对MySpace评论的随机样本的可靠的人工判断，使用单独的数据样本(总共2，600条评论)进行了两次试点练习。这些被用来确定关键的判断问题和适当的规模。虽然测量情绪的方法有很多(Mauss & Robinson，2009；Wiebe，Wilson和Cardie，2005)，人类编码者的主观判断被用作收集足够结果的适当方式。起草并完善了一套编码说明，并构建了一个在线系统来随机选择评论并将其提交给编码人员。试点工作的一个重要成果是，编码者将能量的表达视为积极情绪的表达，除非是在明确的负面背景下。比如“嘿！！!"会被解释为积极的，因为它在没有给出情绪极性线索的上下文中表达能量，所以它会被大多数编码者默认为积极的。相比之下，“窝囊废！！!"会被解释为比“失败者”更消极，因为感叹号与一个消极的词相关联。因此，说明书被修改以明确声明这种表面上中性的能量和积极情绪的融合是允许的。

对于最终的判断，选择了数据集中超过1000条MySpace评论(平均每条评论20个单词和101个字符),以5分制对积极和消极情绪进行判断，如下所示。

[没有积极的情绪或能量]1–2–3–4–5[非常强烈的积极情绪]

[没有负面情绪]1–2–3–4–5[非常强烈的负面情绪]

编码员得到了对每个文本进行编码的口头指示，以及一本解释任务的小册子(由Wiebe等人，2005年激发)，关键指示转载于本文的附录中。这本小册子还包含一个表情符号和首字母缩略词的列表，并附有解释和任务的背景情况，以达到激励的目的。该手册的早期版本包括了与正面和负面情绪判断相关的评论示例，但在试点测试阶段，这些示例在实践中对编码人员几乎没有影响。因此，不使用该组示例，以便可以更真实地评估编码器间的可靠性，而不会出现某些注释与给出的示例过于相似的可能性。

个体对情绪的感知是不同的，部分是因为他们的生活经历，部分是因为性格问题(Barrett，2006)和性别(Stoppard & Gunn Gruchy，1993)。对于系统开发，判断应该对数据中的情绪给出一致的观点，而不是对人口平均感知的估计。结果，使用了一组相同性别(女性)的编码者，并进行初始测试以识别同质子集。最初选择了五名编码员，但后来有两名因给出异常结果而被拒绝:一名给出了比其他人高得多的正面分数，而另一名给出了通常不一致的结果。计算三位编码员对每条评论的平均结果，并四舍五入。这是实验的黄金标准。以下是一些文本和判决的例子。

* 嘿，女巫，你也起床了(分数:+ve: 2，3，1；-ve: 2，2，2)
* omg我儿子和你一样的b日lol(分数:+ve: 4，3，1；-ve: 1，1，1)
* 嘿，你有两个朋友！！(分数:+ve: 2，3，2；-ve: 1，1，1)
* 那个叫卡森的男孩怎么了？(分数:+ve: 1，1，1；-ve: 3，2，1)

表1报告了编码器间的一致程度。此处报告了基本符合率，以便与SentiStrength进行比较。先前的情感判断/注释任务已经获得了较高的编码器间得分，但是没有强度测量，因此具有较少的类别(例如，Wiebe等人，2005)。此外，之前的一篇论文指出，较长(博客)文本的编码者之间的一致程度较高(Gill，Gergle，French和Oberlander，2008年)，这表明在短文本上获得一致将是困难的。对于这种具有多个编码器和不同类别间差异的数据，合适的编码器间可靠性统计类型是Krippendorff的α (Artstein & Poesio，2008；Krippendorff，2004)。使用情绪得分的数字差异作为权重，三个编码器α值对于积极情绪为0.5743，对于消极情绪为0.5634。这些值足够正，表明编码者之间有广泛的一致，但还不够正(例如< 0.67。尽管精确的限制不适用于Krippendorff的带权重的α),以表明编码者一致地测量清晰的底层构造。然而，使用编码者的平均值作为黄金标准似乎仍然是获得情绪强度估计值的合理方法。

表1。编码员之间对1，041条评估意见的一致程度(精确一致、同一类别内一致的百分比、平均百分比误差和皮尔逊相关)。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **比较** | **+ve** | **+ve**  **+/- 1级** | **+ve平均值% diff。** | **+ve corr** | **-ve** | **-ve**  **+/- 1级** | **-ve平均值% diff。** | **-ve corr** |
| 编码器1与2 | 51.0% | 94.3% | .256 | .564 | 67.3% | 94.2% | .208 | .643 |
| 编码器1与3 | 55.7% | 97.8% | .216 | .677 | 76.3% | 95.8% | .149 | .664 |
| 编码器2与3 | 61.4% | 95.2% | .199 | .682 | 68.2% | 93.6% | .206 | .639 |

# SentiStrength情感强度检测算法

SentiStrength情绪检测算法是在用于试点测试的2600个MySpace分类的初始集上开发的。下面列出了SentiStrength的关键要素。

* 算法的核心是情感词强度表。这是298个正面词语和465个负面词语的集合，被分类为正面或负面情绪强度，值从2到5。默认分类是基于开发阶段的人工判断，在训练阶段会自动修改(见下文)。在LIWC之后，一些单词包括通配符(例如，xx\*)匹配任何数量≥2的连续x。一些术语是标准的英语单词，另一些是非标准的，但在MySpace中很常见(例如，luv、xox、lol、哈哈、muah)。情感强度是特定于单词在MySpace中使用的上下文的。例如，“爱”最初被归类为强度4阳性，但由于许多随意使用，如“只是在爱2 ur page中显示”而被降至强度3。一些词明确表达情感，如“爱”或“恨”，但其他词通常被赋予较弱的强度2，与积极或消极的上下文间接相关(如，欣赏、帮助、生日)。SentiStrength算法包括使用一组训练数据来微调情感强度的过程(如下所述)。
* 通过训练算法来修改上述默认手动单词强度，以优化情感单词强度。该算法从预定义列表的基线人类分配的术语强度开始，然后对于每个术语，评估强度增加或减少1是否会增加分类的准确性。任何将整体精度提高至少2的更改都会被保留。最小增量也可以设置为1，这将有过度拟合的风险，而2则有丢失对稀有单词有用的更改的风险。这里选择2是为了使算法运行得更快，因为变化更少，而不是因为任何理论上的原因(事实上，正如结果所示，该算法在测试数据上使用1工作得更好)。该算法随机测试情感列表中的所有单词，并且重复进行，直到所有单词都被检查过，而它们的强度没有改变。
* “小姐”这个词的正负强度分别为2。这是唯一一个既被归类为肯定又被归类为否定的词。它通常用在短语“我想你”中，暗示悲伤和爱。
* 拼写校正算法通过包含重复字母来识别拼写错误的单词的标准拼写。例如，hell 11000将被该算法识别为“hello”。算法(a)自动删除两次以上的重复字母(例如hello-> hello)；(b)对于在英语中很少出现两次的字母，删除出现两次的重复字母(例如，niice -> nice)，以及(c)删除出现两次的字母，如果它们不是标准单词，但如果被删除将形成标准单词(例如，nnice -> nice但不是hoop -> hop nor baaz -> baz)。正式的拼写校正算法(见Pollock & Zamora，1984)被尝试过，但没有被使用，因为它们很少进行校正，并且在姓名和俚语方面存在问题。
* 增强词列表包含增强或降低后续词的情绪的词，无论是积极的还是消极的。每个单词增加1或2的情感强度(例如，非常，极端)或减少1(例如，一些)。
* 否定单词列表包含反转后续情感单词的单词(包括任何前面的助词)。例如，如果“非常快乐”的强度为4，那么“不太快乐”的强度为4。一些否定术语不否定的可能性没有被纳入，因为这在试点数据集中似乎不经常出现。
* 只要有至少两个额外的字母，超过正确拼写所需的重复字母被用来给情绪词增加1的强度。重复字母的使用是在MySpace评论中表达情感或能量的常用手段，但一个重复的字母经常会出现打字错误。
* 具有相关强度(正的或负的2)的表情列表补充了情感单词强度列表(并且表情中包括的标点符号不被进一步处理以用于下面的目的)。
* 任何带有感叹号的句子都被分配了最低2的阳性强度。
* 包括至少一个感叹号的重复标点符号给紧接在前面的情感词(或句子)增加1的强度。
* 负面情绪在提问中被忽略。比如问题“你生气了吗？”将被归类为不包含情绪，尽管存在“愤怒”一词。这不适用于积极情绪，因为许多疑问句似乎包含温和的积极情绪。特别是像“怎么了？”通常被归类为包含轻度积极情绪(强度2)。

上述因素分别应用于每个句子，句子被分配了最积极和最消极的情绪。每条总体评论都被分配了最积极的句子情绪和最消极的句子情绪。句子被注释中换行符或者除了表情符号以外的标点符号分开。

SentiStrength增加了一些额外的修改，但随后在额外的测试后被拒绝，或被发现不切实际。

* 除了在最初的2600条开发意见中发现的几个常见例子之外，短语识别没有被广泛使用。虽然惯用短语很常见，但它们的多样性使得系统地识别它们似乎不太实际。未来的工作可能会识别像“非常”和“很多”这样的增强短语，并使用短语识别来区分“爱”这个词的弱用法和强用法，如“我爱你”。
* 由于高度不规范的语法导致的问题，语义消歧没有用于歧义词。这可能会改进算法，但需要相当大的计算量。例如，单词“rock”有时是非常积极的(例如，you rock！！！)有时是中性的(例如，你听摇滚乐吗？).

# 实验

SentiStrength在一组1041条MySpace评论上进行测试，这些评论不同于开发阶段使用的评论，由三个人进行分类(见表1)，平均值被用作黄金标准。使用了10倍交叉验证方法。将结果与随机分配和基线多数类分类(积极情绪为2，消极情绪为1)进行比较。SentiStrength还与Weka (Witten & Frank，2005)中的一系列标准机器学习分类算法进行了比较，使用情感单词列表中每个单词的频率作为特征集。用于比较的扩展特征集包括长度为1-3的n元语法，由从文本中提取的所有术语组成，包括表情符号、拼写纠正的单词(在适当的情况下)、重复的标点符号、问号和感叹号(例如，一个特征是3元语法:“love-u-！”)以及每个注释中1、2和3个字母的总数的计数。这个扩展的特性集包含了SentiStrength使用的大部分文本元素。

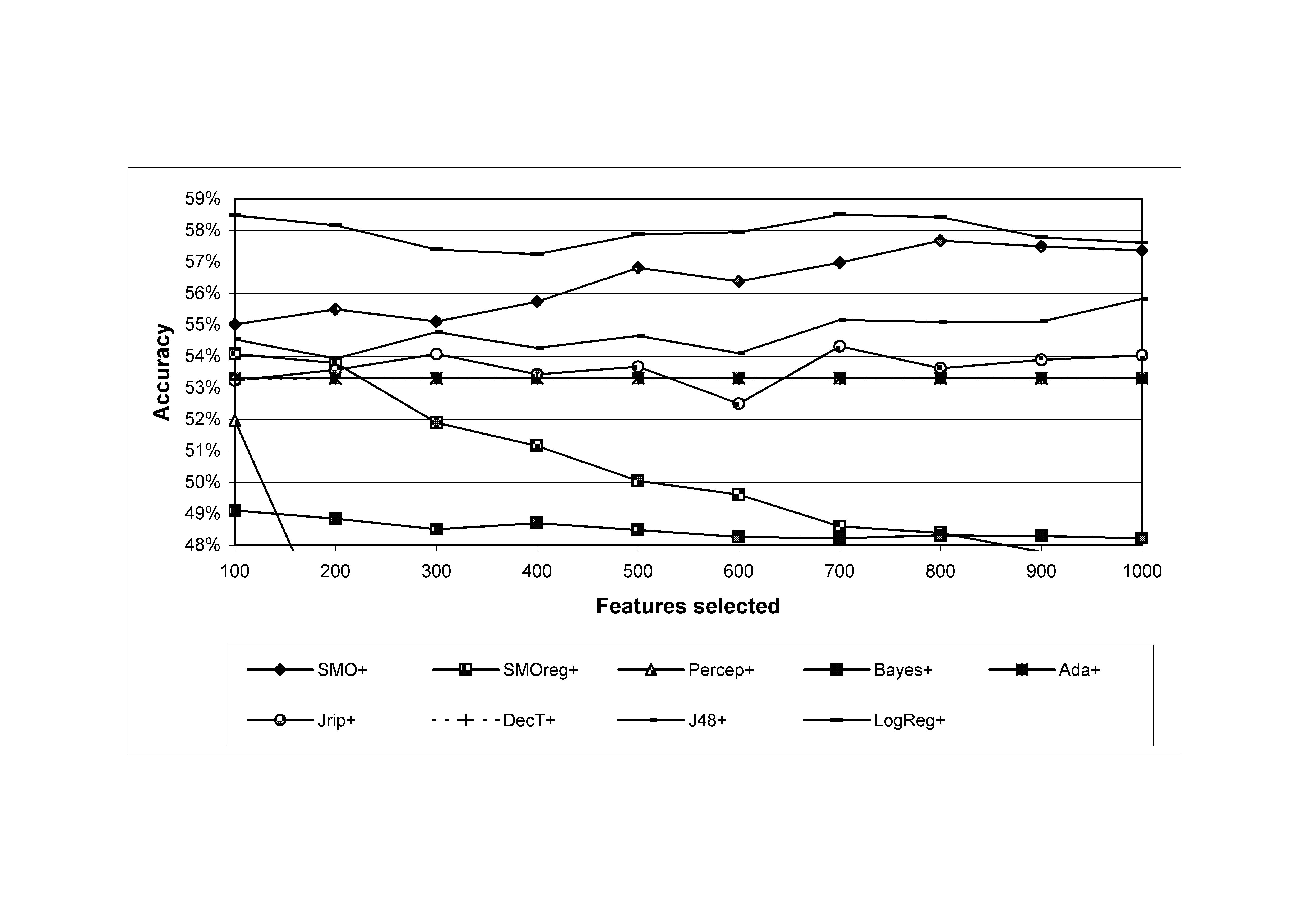
第二个测试比较了不同的功能集，以查看替代的较小功能集是否可以为机器学习提供更好的结果，并发现哪些功能最有用。

第三个测试使用包含的特征缩减(详见下文)。

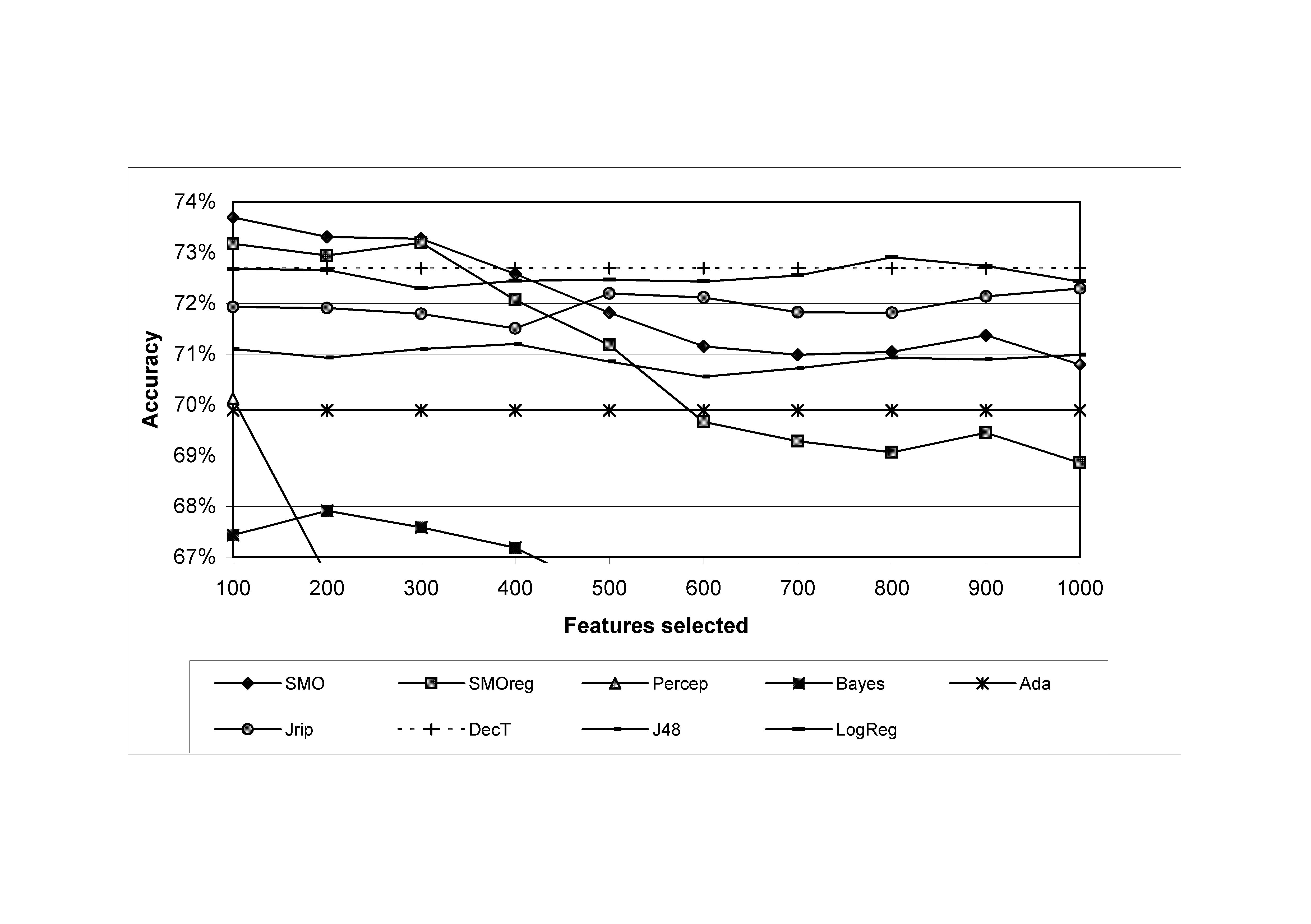
第四个测试比较了SentiStrength的不同变体，以查看算法的哪些方面最强大。

## 与机器学习、扩展功能集的比较

图1和图2显示了各种机器学习算法在具有不同特征集大小的1，041条MySpace评论上的性能，这些特征集大小是使用来自信息增益度量的顶级特征选择的。特征选择改善了所有方法的结果，只有一个小的例外(正面情绪的朴素贝叶斯:52.0%没有特征选择，平均超过4个10倍交叉验证)。对于每种方法，表2报告了使用每种方法的最佳特征集大小与SentiStrength的比较。



图一。针对使用扩展特征集的不同分类器的特征集大小的正面情感分类准确度；平均超过4个分类。



图二。针对使用扩展特征集的不同分类器的特征集大小的负面情感分类准确度；平均超过4个分类。

从表2中可以看出，使用具有最佳特征数量的扩展特征集的机器学习分类器，如通过信息增益所选择的，明显不如SentiStrength准确。SentiStrength还具有与黄金标准最高的相关性、最低的平均百分比误差和最高的一级精度。因此，它始终比其他算法表现更好(至少2.1%)。然而，SentiStrength的准确度为60.6%。这类似于人类编码者之间的一致程度(表1)，表明在非正式短文本中的积极情绪强度检测是一项固有的困难任务。

表二。使用扩展的特征集和10重交叉验证(正面情感强度性能的降序)对1，041条评论的正面情感强度检测的各种算法的性能。除了SentiStrength，结果是不同随机测试/训练分割的4次运行的平均值，并且是从图1中选择的最佳特征数。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **算法** | **最佳的**  **特征** | **准确(性)** | **准确(性)**  **+/- 1**  **班级** | **科尔。** | **平均百分比**  **绝对的**  **错误** |
| **感知强度**  (标准配置，30次运行) | - | 60.6% | 96.9% | .599 | 22.0% |
| 简单逻辑回归 | 700 | **58.5%** | 96.1% | **.557** | **23.2%** |
| SVM (SMO) | 800 | **57.6%** | **95.4%** | **.538** | **24.4%** |
| J48分类树 | 700 | **55.2%** | **95.9%** | .548 | 24.7% |
| JRip基于规则的分类器 | 700 | **54.3%** | 96.4% | **.476** | **28.2%** |
| SVM回归(SMO) | 100 | **54.1%** | 97.3% | **.469** | **28.2%** |
| adaboost算法 | 100 | **53.3%** | **97.5%** | **.464** | **28.5%** |
| 决策表 | 200 | **53.3%** | 96.7% | **.431** | **28.2%** |
| 多层感知器 | 100 | **50.0%** | *94.1%* | **.422** | **30.2%** |
| 朴素贝叶斯 | 100 | **49.1%** | **91.4%** | **.567** | **27.5%** |
| 基线 | - | **47.3%** | **94.0%** | **-** | **31.2%** |
| 随意 | - | **19.8%** | **56.9%** | **.016** | **82.5%** |

与SentiStrength相比，Bold=sig为0.01，italic=sig为0.05。

对于负面情绪强度，大多数方法给出了非常相似的结果，一些方法给出了比情绪强度更好的结果。虽然SentiStrength的准确率为72.8%，仅比基线高2.9%，但其他几种方法的准确率相似，而SVM的准确率明显更高。如果最多允许一个类别的误差，SentiStrength显然是这些方法中最准确的，并且与人类编码器结果具有明显最高的相关性。请注意，理论上，任何方法都不应比基线差，但这可能是由于对训练集而不是评估集进行了优化。总的来说，感觉强度似乎不擅长识别负面情绪，但对于本文分析的短文来说，这是一项艰巨的任务。还要注意，随机类别的平均百分比绝对误差超过100%,因为“1”是负面情绪的正确类别。

表3。使用扩展的特征集和10重交叉验证(正面情感强度性能的降序)对1，041条评论的负面情感强度检测的各种算法的性能。除了SentiStrength之外，结果是4次运行的平均值，并且是从图2中选择的最佳特征数。

与SentiStrength相比，Bold=sig为0.01，italic=sig为0.05。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **算法** | **最佳的**  **特征** | **准确(性)** | **准确(性)**  **+/- 1**  **班级** | **科尔。** | **平均百分比**  **绝对的**  **错误** |
| SVM (SMO) | 100 | *73.5%* | **92.7%** | **.421** | **16.5%** |
| SVM回归(SMO) | 300 | 73.2% | **91.9%** | **.363** | 17.6% |
| 简单逻辑回归 | 800 | 72.9% | **92.2%** | **.364** | 17.3% |
| **感知强度**  (标准配置，30次运行) | - | 72.8% | 95.1% | .564 | 18.3% |
| 决策表 | 100 | 72.7% | **92.1%** | **.346** | **17.0%** |
| JRip基于规则的分类器 | 500 | 72.2% | **91.5%** | **.309** | 17.3% |
| J48分类树 | 400 | 71.1% | **91.6%** | **.235** | 18.8% |
| 多层感知器 | 100 | 70.1% | **92.5%** | **.346** | 20.0% |
| adaboost算法 | 100 | **69.9%** | **90.6%** | **-** | **16.8%** |
| 基线 | - | **69.9%** | **90.6%** | **-** | **16.8%** |
| 朴素贝叶斯 | 200 | **68.0%** | **89.8%** | **.311** | **27.3%** |
| 随意 | - | **20.5%** | **46.0%** | **.010** | **157.7%** |

本文的其余部分只关注积极情绪，因为消极情绪的结果并不显著。

## 机器学习的特征集比较–积极情绪强度

图3和图4比较了使用不同特征集和两种最佳执行算法对正面情感强度检测的影响。特征集是:1-3个字母；1-3克带表情符号；1-3-克带标点符号；有拼写错误的1-3个单词(即，除了拼写校正后的单词之外，还包括拼写校正前的单词，如果不同的话)；有表情符号、标点符号和拼写错误的1-3个字母；1-3克带情感术语；和1克。基本的袋或词方法(1-grams)表现很差，对于逻辑回归总是最差的特征集，对于SVM总是最差或在最差的几个特征集中。对于SVM来说，最好的结果是用通过情感项增强的基本1-3-grams实现的，尽管大多数时候(即，对于500-1000个特征)扩展特征集(标记为“以上所有”并且与以上结果中使用的相同)表现最好，可能主要是由于标点符号成分，因为这种增强对于700-1000个特征表现第二好。

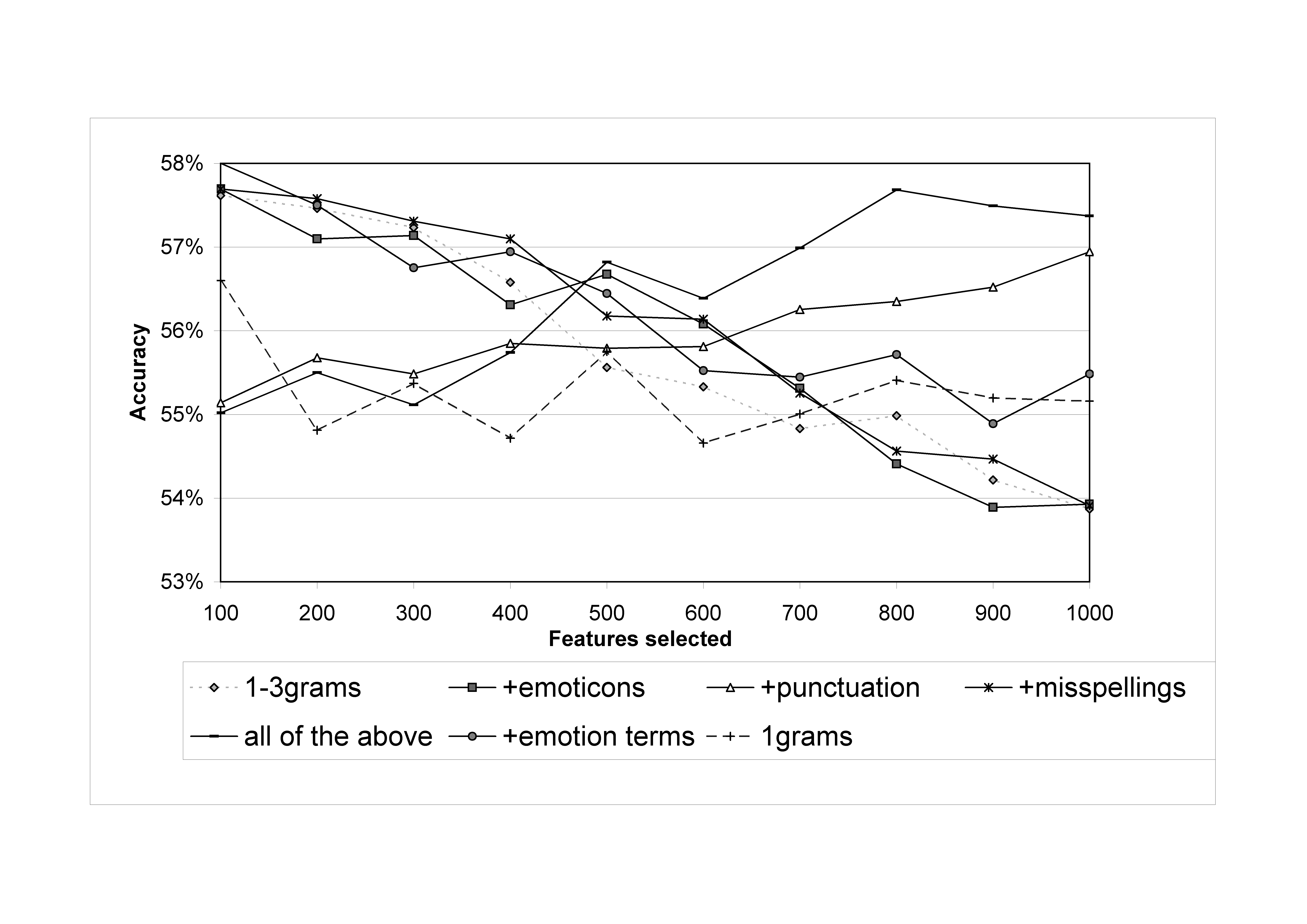


图3。SVM (SMO)对不同特征集类型的特征集大小的正面情感分类准确度；平均超过4个分类。

图4表明，除了基本的单词包之外，逻辑回归的特征集之间的差异没有SVM的明显，但是表现最好的组合仍然是1-3克加情感术语。对于较大的特征集，组合的特征集表现最好，可能是由于标点符号和情感项。

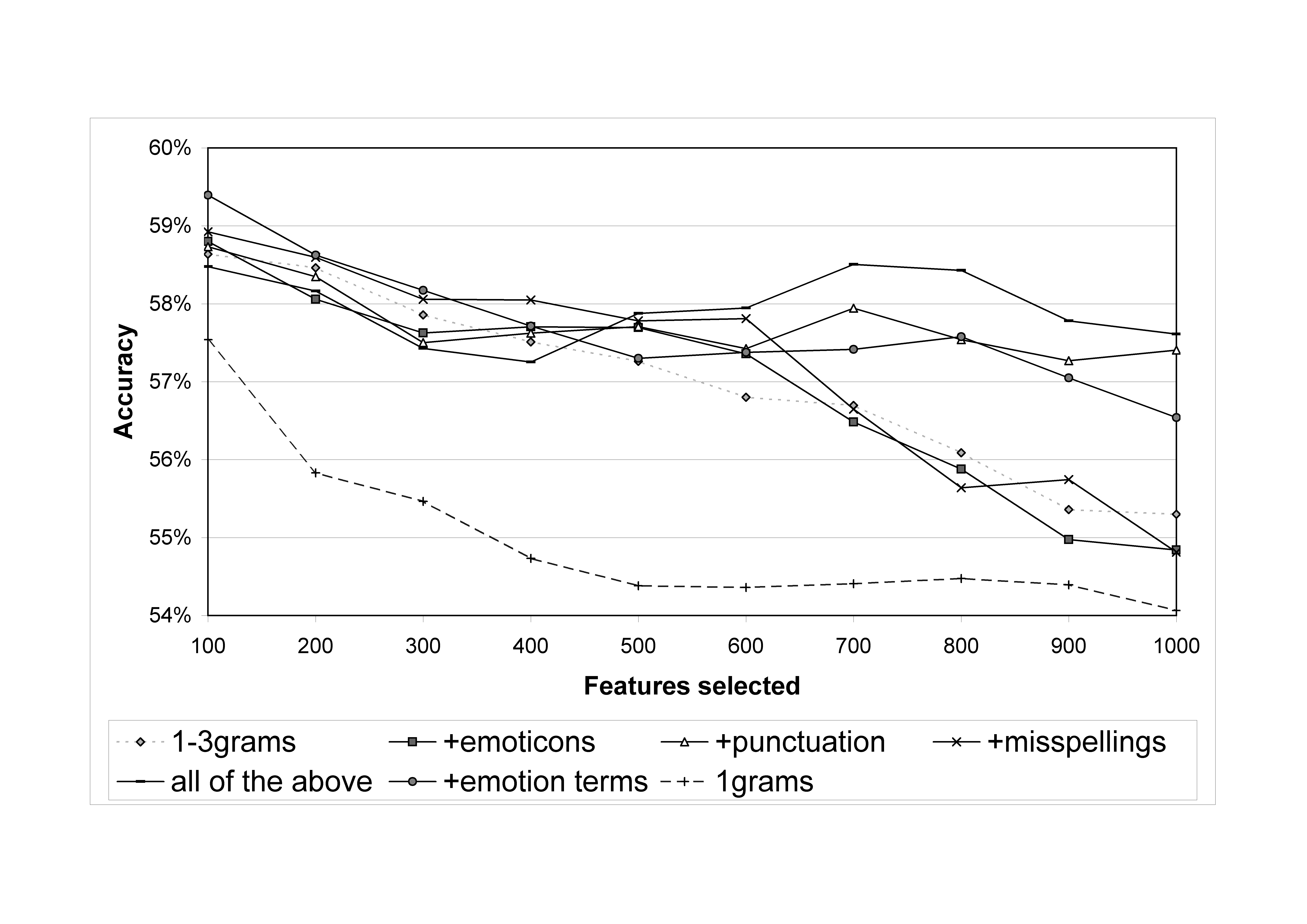


图4。针对不同特征集类型的特征集大小的逻辑回归正情感分类准确度；平均超过4个分类。

将二元模型和三元模型与一元模型结合使用的一个潜在缺点是存在一些冗余。例如，三元组“我爱你”也将匹配双元组“我爱”和“爱你”，以及单元组“我”、“爱”和“你”。作为响应，包容是一种特征选择方法，它消除了二元模型和三元模型，这些模型看起来是冗余的，因为除了它们的组成元模型(和二元模型的三元模型)之外，没有给出额外的信息。这种方法在这里是合适的。包容是通过一个逻辑扩展来应用的:如果适当的度量被匹配，像happ\*这样的词模式可以消除匹配的词(例如，happy，在这种情况下是happy)。图5和图6显示了其表现最佳的两种机器学习算法的包含结果:SVM和逻辑回归。如两幅图所示，包容与特征缩减结合起来表现最好。对于其他算法，包含将Jrip的性能提高了0.4% (α = 0.005，100个特征)，将SVM回归的性能提高了1.1% (α = 0.02，100个特征)，将多层感知器的性能提高了1.0% (α = 0.02，100个特征)，将决策表的性能提高了1.0% (α = 0.005，900个特征)，但没有提高J48、AdaBoost和朴素贝叶斯的性能。

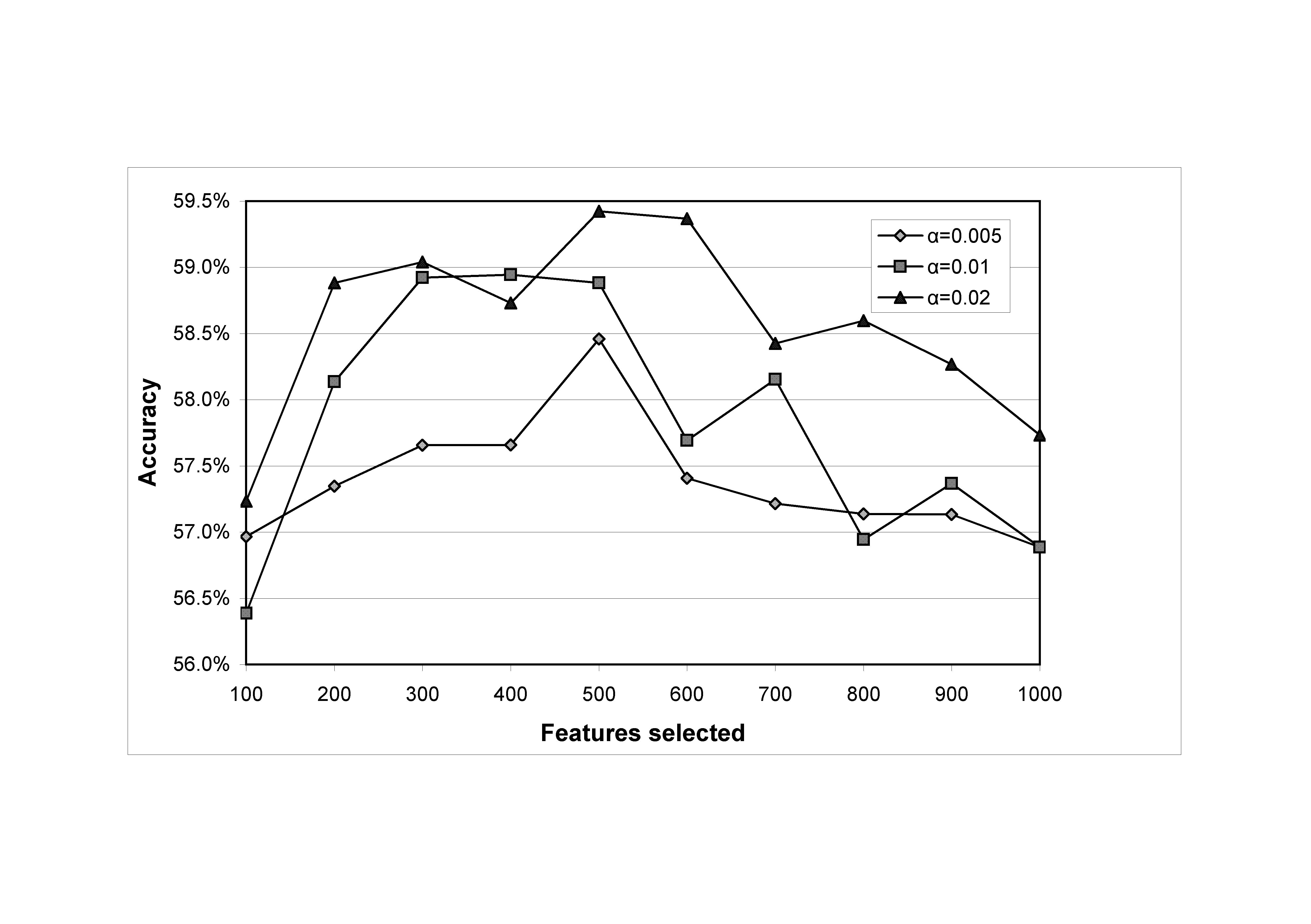


图5。SVM(SMO)正情感分类准确度相对于特征集大小，用于包含各种α值；平均超过5个分类。

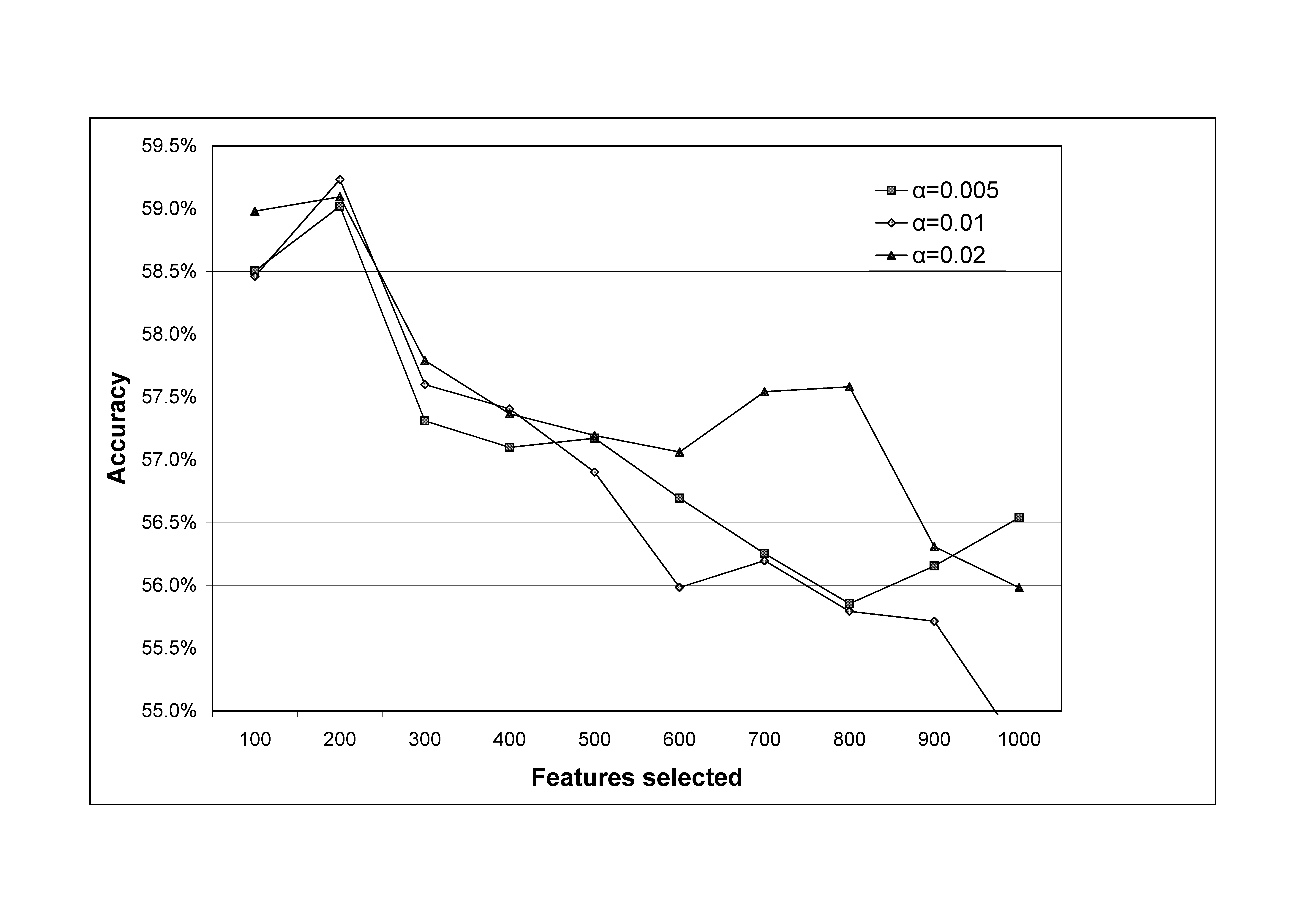


图6。逻辑回归正情感分类准确度与特征集大小的关系，用于包含各种α值；平均超过5个分类。

从图5中可以看出，在扩展特征集上，包含的SVM比没有包含的SVM高出1.8%，在所有其他特征集上(α= 0.02500个特征)都超过了SVM。尽管如此，其准确性低于SentiStrength标准版本，尽管差异在统计学上不显著(准确性= 59.42%，准确性+/-1 = 96.60%，相关性= 0.5822，平均绝对误差= 22.65%；只有平均绝对误差差与SentiStrength标准配置相比具有统计显著性)。从图6中可以看出，包含的逻辑回归优于不包含扩展特征集的逻辑回归，低0.7% (α = 0.01，200个特征)。然而，在添加了情感项的情况下，它的表现不如1-3克好(图4)，但由于进行了大量的比较，这可能是一种统计异常。逻辑回归的表现不如标准的SentiStrength，但差异也不显著(准确度= 59.23%，准确度+/-1 = 95.79%，相关性= 0.5820，平均绝对误差= 22.57%；除了准确性之外，所有都与SentiStrength标准配置有统计学上的显著差异)。就α值而言，对于该数据集，0.02的表现几乎一致优于其他值。

注意，尽管SentiStrength在统计上并不显著优于使用包容的最优SVM和逻辑回归模型，但是表4中的SentiStrength的最优变化，具有一个简单的修改(训练仅需要增加1来改变单词强度)，在统计上在所有方面都显著优于SVM，并且在统计上在所有方面都显著优于逻辑回归，除了+/-1内的准确度。

## SentiStrength版本的比较

表4和表5报告了感觉强度不同变化的比较。大多数变量对结果几乎没有影响——除了最后两个选项，每个变量最多占算法性能的0.8%。这些差异小到足以归因于所使用的语料库，因此该表没有提供令人信服的证据来证明任何变化比标准方法更好或更差。当删除所有选项(但不改变平均方法)时，累积效应更显著，但是性能降低了3.4%。也许使用非标准特性的注释倾向于使用多个非标准特性，所以如果一个特殊规则被忽略，那么这通常会被其他特殊规则所补偿。

与表2和表3相比，SentiStrength的主要优势在于其适应各种非正式文本变化的规则的综合效果，以及使用术语优势列表和识别任何评论中最强的正面和负面术语的整体方法。在这种情况下，表2中的通用分类算法似乎比SentiStrength的效率至少低2.1%，主要是因为1-3克的方法不够灵活，无法处理非标准的MySpace语言(大约3.4%归因于这一原因)。此外，他们似乎无法利用足够大的训练集来学习有效的术语优势，而更大的训练集可以看到其中一些更接近SentiStrength的表现。最后，请注意，SentiStrength的变化虽然明显改善了它，但并不是更健壮:当所有这些组合在一起形成新版本的SentiStrength时，它具有与标准配置完全相同的准确性(60.64%正确，97.07% +/- 1类，0.6071相关性，21.62%平均%误差)。

表4。几种算法变体的正面情感表现的比较:对1，041条分类评论进行30次10倍交叉验证的平均。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类型 | %正确 | **+/- 1级** | **科尔。** | **平均%误差。**  **(pred-act)/act** |
| SentiStrength标准算法(但是训练只需要增加1来改变单词强度) | **61.03%** | **96.68%** | .5983 | **21.66%** |
| 不用于转换后续情绪的否定词(例如，不开心) | 60.87% | **97.50%** | **.6206** | **21.28%** |
| 多个连续的积极词汇不能用作情绪助推器 | 60.70% | 96.88% | .5962 | 21.97% |
| 表情被忽略 | 60.68% | 96.87% | .5977 | 21.95% |
| 忽略助词(例如，非常) | 60.68% | **96.80%** | .5970 | **22.14%** |
| **SentiStrength标准算法** | 60.64% | 96.90% | .5986 | 21.96% |
| 感叹号的强度不是2 | 60.51% | **96.62%** | **.6035** | **21.47%** |
| 自动拼写纠正已禁用 | *60.39%* | 96.88% | *.5961* | 22.05% |
| 额外的多个字母不能用作情绪助推器 | **60.21%** | **96.81%** | **.5952** | **22.16%** |
| 术语“错过”没有被赋予+2的强度 | 60.45% | **96.77%** | **.5953** | **22.16%** |
| 习语查找表已禁用 | 60.52% | 96.88% | **.6054** | **21.62%** |
| 带强调的中性词不算积极情绪 | **60.13%** | **96.79%** | .5966 | 21.90% |
| 感受所有上述变化带来的力量 | **57.44%** | **96.07%** | **.6073** | 21.91% |
| 句子情感是所有术语情感的平均值(而不是最大值) | **42.40%** | **88.54%** | **.4065** | **29.27%** |
| 文本情感是所有句子情感的平均值(而不是最大值) | **39.13%** | **86.96%** | **.3293** | **33.19%** |

\*粗体= p = 0.01时显著，斜体=sig。与标准算法相比，p=0.05。

表5。几种算法变体的负面情绪表现的比较:对1，041条分类评论进行30次10倍交叉验证的平均。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类型 | %正确 | **+/- 1级** | **科尔。** | **平均%误差。**  **(pred-act)/act** |
| 问题中的负面情绪是不容忽视的 | **73.56%** | **95.14%** | **.5921** | **18.11%** |
| SentiStrength标准算法(但是训练只需要增加1来改变单词强度) | 72.95% | **94.86%** | .5651 | **18.16%** |
| 不用于转换后续情绪的否定词(例如，不开心) | 72.84% | **94.79%** | **.5706** | 18.35% |
| **SentiStrength标准算法** | 72.83% | 95.07% | .5644 | 18.27% |
| 多个连续的否定词不能用作情绪助推器 | 72.81% | 95.08% | .5653 | 18.29% |
| 表情被忽略 | 72.80% | **94.97%** | .5614 | 18.28% |
| 除平均外，用此表中的所有变化感知强度 | 72.76% | **94.59%** | .5668 | **19.07%** |
| 习语查找表已禁用 | *72.73%* | *95.03%* | **.5556** | **18.63%** |
| 额外的多个字母不能用作情绪助推器 | 72.72% | 95.04% | .5627 | **18.40%** |
| 文本情感是所有句子情感的平均值(而不是最大值) | 72.66% | **95.83%** | **.5486** | **16.81%** |
| 自动拼写纠正已禁用 | **72.64%** | 95.07% | **.5586** | **18.62%** |
| 忽略助词(例如，非常) | **72.35%** | *95.03%* | **.5559** | **18.50%** |
| 句子情感是所有术语情感的平均值(而不是最大值) | **72.17%** | **95.35%** | **.4980** | **16.82%** |

\*与标准算法相比，粗体= p = 0.01时显著，斜体= p = 0.05时sig.t。

表5示出了用于负面情绪强度检测的SentiStrength的不同变体的性能变化非常小:性能与标准配置的最大差异为0.83%。然而，这表明问题中的消极情绪(例如，“你恨托尼吗？”)今后不应被忽视。

# 讨论和结论

回想一下，这篇论文的主要新颖贡献是:优化情感术语权重的机器学习方法；从文本中的非标准拼写中提取情感的方法；以及相关的拼写纠正方法。SentiStrength能够在60.6%的时间里用非正式的MySpace语言在1到5的范围内识别积极情绪的强度，显著高于性能高达58.5%的最佳标准机器学习方法，这与之前的4类意见强度分类任务一致(Wilson等人，2006年)。SentiStrength的标准版本也优于标准的机器学习方法，当它们的性能通过使用包含和信息增益特征减少而得到改善(或在某些情况下没有改善)时，但这种差异在统计上并不显著。然而，稍加修改的SentiStrength版本在统计上明显优于改进的机器学习方法。这很好地证明了SentiStrength对于正面情感强度检测的功效，给定了不同算法和与之比较的参数的范围(9个算法x 11个特征集大小，x 7个特征集类型= 693个变化，加上9个算法x 10个特征集大小x 3个α值= 270个包容变化)，这给了较低性能的算法一个通过机会胜过SentiStrength的合理的统计机会，但没有一个算法这样做。

SentiStrength相对成功的主要原因似乎是解码非标准拼写的程序和增强单词强度的方法，这是其性能的主要原因。在没有这些因素的情况下，仅基于情感相关词汇词典的SentiStrength变体及其57.5%的估计强度仅比最成功的机器学习方法在1-3克的扩展集上好1.3%。相比之下，SentiStrength能够识别的负面情绪比基线好不了多少(1.8%)，这可能是因为在表达负面评论时的创造力，或者是因为当一个类别占主导地位时，很难明显高于基线(Artstein & Poesio，2008；Krippendorff，2004)。在像MySpace评论这样的非正式文本语言中，正面和负面情绪检测似乎都具有挑战性，这是因为几个因素:语言创造性、没有情绪承载词的情绪表达，以及人类编码者解释之间的差异，这意味着对于大多数评论来说没有真正正确的分类。

鉴于在生成正面情感强度检测的算法方面的成功以及正面情感在MySpace评论中的优势，看来未来的研究可以应用情感强度检测技术来大规模地自动识别和分类非正式网络交流环境中的正面情感。此外，有许多情感分析的商业应用，其中一些使用从聊天室或移动电话文本消息生成的非正式计算机文本，并且该算法表明，即使在这些短消息中，也可以估计积极情感的强度。

就今后的工作而言，下一个合乎逻辑的步骤是试图通过语言处理提高系统的性能，尽管所分析的简短非正式文本信息的语法很差。先前的工作表明，这种方法是有前途的，特别是通过依赖树(Wilson等人，2009)，并且，给定足够大的训练样本，即使在质量差的文本中也有可能改进(Gamon，2004)。

# 附录:编码器说明(摘录)

根据评论表达积极情绪或能量的程度进行编码。兴奋、热情或精力在这里应该算作积极情绪。如果你认为标点符号以任何方式强调了积极的情绪或能量，那么在你的评分中包括这一点。积极情绪或能量的尺度是:

[没有积极的情绪或能量]1–2–3–4–5[非常强烈的积极情绪]

* 如果评论不包含积极的情绪或能量，请给1分。
* 如果评论包含非常强烈的正面情绪，请给5分。
* 如果评论包含一些积极的情绪，但不是非常强烈的积极情绪，请分配一个介于2和4之间的数字。用你的判断来判断积极情绪的确切强度。

对每条评论进行编码，标明它表达负面情绪或消极的程度。如果你认为标点符号以任何方式强调了负面情绪，那么在你的评分中包括这一点。负面情绪的等级是:

[没有负面情绪]1–2–3–4–5[非常强烈的负面情绪]

* 如果评论中完全没有负面情绪，请给1分。
* 如果评论包含非常强烈的负面情绪，请给5分。
* 如果评论包含一些负面情绪，但不是非常强烈的负面情绪，请分配一个介于2和4之间的数字。用你的判断来判断负面情绪的确切强度。

做出判断时，请尽可能与您之前的决定保持一致。此外，请解释个人评论中出现的情绪，并忽略所有其他评论。

# 参考

Abbasi，a .，Chen，h .，和Salem，A. (2008年)。多语言中的情感分析:网络论坛中观点分类的特征选择。ACM信息系统汇刊，26(3)，12:11-12.34。

阿巴西，陈，h，汤姆斯，s .，，傅，T. (2008)。使用相关集成的网络论坛和博客的影响分析。IEEE知识与数据工程汇刊，20(9)，1168-1180。

Agerri，r .，& García-Serrano，A. (2010年)。Q-WordNet:从WordNet感官中提取极性。第七届国际语言资源与评估会议记录，2010年5月25日检索自:http://www . lrec-conf . org/proceedings/lrec 2010/pdf/2695 \_ paper . pdf。

Argamon，s .，Whitelaw，c .，Chase，p .，Hota，S. R .，Garg，n .，和Levitan，S. (2007年)。基于功能词汇特征的文体文本分类。美国信息科学与技术学会杂志，58(6)，802-822。

Artstein，r .，& Poesio，M. (2008年)。计算语言学编码器间协议。计算语言学杂志，34(4)，555-596。

Baccianella，s .，Esuli，a .，和Sebastiani，F. (2010年)。SentiWordNet 3.0:用于情感分析和观点挖掘的增强型词汇资源。第七届国际语言资源与评估会议记录，2010年5月25日检索自:http://www . lrec-conf . org/proceedings/lrec 2010/pdf/2769 \_ paper . pdf。

Balahur，a .，Kozareva，z .，和Montoyo，A. (2009年)。确定政治辩论中所表达观点的极性和来源。计算机科学课堂讲稿，5449，468-480。

Balahur，a .，Steinberger，r .，Kabadjov，m .，Zavarella，v .，Goot，E. v. d .，Halkia，m .，等人(2010年)。新闻中的情感分析。第七届国际语言资源与评估会议记录，2010年5月25日检索自:http://www . lrec-conf . org/proceedings/lrec 2010/pdf/2909 \_ paper . pdf。

拜伦，N. S. (2003年)。互联网的语言。在A. Farghali(编辑。)，斯坦福语言工程师手册(第59-127页)。斯坦福:CSLI出版公司。

巴雷特，L. F. (2006年)。价是情感生活的基本组成部分。个性研究杂志，40(1)，35-55页。

博伊德博士(2008年)。断章取义:网络公众中的美国青少年社会性。加州大学伯克利分校。

博伊德博士(2008年)。为什么青少年(心脏)社交网站:网络公众在青少年社交生活中的作用。在d .白金汉(编辑。)、青年、身份和数字媒体(第119-142页)。麻省剑桥:麻省理工学院出版社。

布里尔，E. (1992年)。一个简单的基于规则的词性标注器。第三届应用自然语言处理会议论文集，152-155。

f . r .肖马丁(2007年)。UPAR7:基于知识的标题情感标注系统。《第四届语义评估国际研讨会论文集》(SemEval-2007)(第422-425页)。纽约州纽约市:美国计算机学会。

蔡宇略和卡迪(2008)。使用合成语义作为结构推理的学习用于子句子情感分析。自然语言处理经验方法会议录，793-801。

科恩，W. (1995年)。快速有效的规则归纳。第十二届机器学习国际会议论文集，115–123。

科尼利厄斯(1996年)。情感科学。新泽西州上马鞍河:普伦蒂斯霍尔。

克里斯特尔博士(2006年)。语言与互联网(第二版。).英国剑桥:剑桥大学出版社。

陈明辉，陈小明(2001)。雅虎！对于亚马逊:从股票留言板中提取市场情绪。亚太金融协会年会(APFA)会议记录，泰国曼谷，7月22-25日，2009年7月17日检索自:http://sentiment.technicalanalysis.org.uk/DaCh.pdf.

Denecke，k .，& Nejdl，W. (2009年)。医疗社交媒体数据有多大价值？医学网站的内容分析。信息科学，179(12)，1870-1880。

Derks，Bos，A. E. R .，& von Grumbkow，J. (2008年)。表情符号和在线消息解释。社会科学计算机评论，26(3)，379-388页。

Derks、a . h . Fischer和a . e . r . Bos(2008年)。情绪在计算机媒介交流中的作用:综述。人类行为中的计算机，24(3)，766–785页。

迪纳和埃蒙斯(1984年)。积极情感和消极情感的独立性。人格与社会心理学杂志，47卷5期，1105-1117页。

埃克曼博士(1992年)。基本情感的论证。认知与情绪，6(3/4)，169-200。

Esuli，a .，& Sebastiani，F. (2006年)。SENTIWORDNET:一个公众可用的观点挖掘词汇资源。语言资源与评估(LREC) 2006年会议录，2009年7月28日检索自:http://TCC . fbk . eu/projects/onto text/Publications/lrec 2006-e suli-sebastiani . pdf

福克斯(2008年)。情感科学。巴辛斯托克:帕尔格雷夫·麦克米伦。

Fullwood和o . I . Martino(2007年)。表情符号和印象形成。流行文化中的视觉，19卷7期，第4-14页。

Gamon，M. (2004年)。顾客反馈数据的情感分类:噪声数据、大特征向量和语言分析的作用。第20届计算语言学国际会议论文集，第841期。

Gamon，m .，Aue，a .，Corston-Oliver，s .，和Ringger，E. (2005年)。脉动:从自由文本中挖掘客户意见(IDA 2005)。计算机科学课堂讲稿，3646，121-132。

吉尔，A. J .，格尔，d .，弗伦奇，R. M .，，奥伯兰德，J. (2008)。短博客文本的情感评级。第26届SIGCHI计算系统中人的因素年会论文集(第1121-1124页)。纽约州纽约市:美国计算机学会。

格林特和埃尔德里奇(2003年)。Wan2tlk？日常短信。池2003，441-448。

汉考克，J. T .，吉，k .，恰乔，k .，，林，J. M.-H. (2008)。我难过你难过:CMC中的情绪传染。ACM 2008年计算机支持的合作工作会议录，295-298。

霍普金斯大学法学博士和金(2010年)。社会科学的自动非参数内容分析方法。美国政治科学杂志，54卷1期，第229-247页。

黄延平、吴廷泰、刘春林(2007)。在web 2.0中寻找遗书-初步发现。第九届Ieee多媒体国际研讨会论文集(第517-521页)。洛斯阿拉米斯:IEEE。

胡珀特，F. A .，&惠廷顿，J. E. (2003)。积极和消极幸福独立性的证据:生活质量评估的含义。英国健康心理学杂志，8(1)，107-122。

Kaji和m . Kitsuregawa(2007年)。从大量HTML文档中构建情感分析词典。《2007年自然语言处理和计算自然语言学习经验方法联合会议论文集》(第1075-1083页，1028年7月检索自:http://www . ACL web . org/prophet/D/D 1007/D 1007-1115 . pdf)。

k . krippendorff(2004年)。内容分析:方法论介绍。千橡，加州:鼠尾草。

Kukich，K. (1992年)。自动更正文本中单词的技术。美国计算机学会计算调查，24(4)，377-439。

刘，李柏曼，h .，，塞尔克，T. (2003)。使用真实世界知识的文本情感感知模型。2003年智能用户界面国际会议论文集，IUI 2003，125-132。

Mauss，I. B .，& Robinson，M. D. (2009)。情绪测量:综述。认知与情绪，23(2)，209-237。

米什内，G. (2005年)。博客文章中的情绪分类实验。文体——2005年SIGIR第一届信息获取文本文体分析研讨会。

g .米什内和m .德里基(2006年)。使用博客帖子捕捉全球情绪水平。《AAAI之春网络日志分析计算方法研讨会论文集》(AAAI-CAAW)(第145-152页)。门洛帕克，加州:AAAI出版社。

纳尔迪，文学学士(2005年)。超越带宽:人际交流中联系的维度。计算机支持的合作工作，14(1)，91-130。

Neviarouskaya，a .，Prendinger，h .，& Ishizuka，M. (2007年)。用于社交和表达性在线交流的文本情感感知。计算机科学课堂讲稿，4738，218-229。

Ng，v .，Dasgupta，s .，，Arifin，S. M. N. (2006年)。检验语言学知识来源在评论的自动识别和分类中的作用。COLING/ACL 2006年主要会议录，611-618。

庞，b .，，李，L. (2004)。情感教育:使用基于最小割的主观性摘要的情感分析。在2004年ACL会议录中(第271-278页)。纽约:ACL出版社。

庞，b .，，李，L. (2005)。根据等级量表利用阶级关系进行情感分类。美国律师协会第43届年会会议录，115-124。

庞，b .，，李，L. (2008)。观点挖掘和情感分析。信息检索的基础和趋势，1(1-2)，1-135。

Pennebaker，j .，Mehl，m .，，Niederhoffer，K. (2003)。自然语言使用的心理学方面:我们的话语，我们的自我。心理学年度评论，54，547-577。

彭尼贝克，J. W .，梅恩，t .，，弗朗西斯，M. E. (1997)。适应性丧亲的语言学预测。人格与社会心理学杂志，72卷4期，863-871页。

波洛克，J. J .，&萨莫拉，A. (1984)。科学和学术文本中的自动拼写校正。美国计算机协会的来文，27(4)，358-368页。

普拉博沃，r .，& Thelwall，M. (2009年)。情感分析:一种综合方法。信息计量学杂志，3(1)，143-157。

里德，J. (2005年)。使用表情符号减少情感分类机器学习技术中的依赖性。ACL 2005年学生研究研讨会会议录，43-48。

里洛夫，e .，帕特瓦丹，s .，，威伯，J. (2006年)。用于观点分析的特征包容。自然语言处理经验方法会议录，440-448。

e .里洛夫和j .韦贝(2003年)。学习主观表达的抽取模式。2003年自然语言处理经验方法会议记录(EMNLP-03)，2010年4月11日从http://www.cs.utah.edu/~riloff/pdfs/emnlp2003.pdf.检索

拉塞尔，J. A. (1979年)。情感空间是两极的。人格与社会心理学杂志，37卷3期，345-356页。

沙皮雷和辛格(2000年)。BoosTexter:一个基于boosting的文本分类系统。机器学习，39(2/3)，135-168。

肖特，J. C .，&帕尔默，T. B. (2008年)。DICTION在战略管理内容分析研究中的应用。组织研究方法，11卷4期，第727-752页。

斯奈德和巴兹莱(2007年)。使用好悲伤算法的多方面排名。NAACL HLT会议录。

斯通，P. J .，邓菲特区，史密斯，硕士和奥吉尔维，D. M. (1966)。内容分析的计算机方法。麻省剑桥:麻省理工学院出版社。

斯托帕德，J. M .，&甘格鲁希，C. D. (1993)。性别、背景和积极情绪的表达。个性与社会心理学通报，19(2)，143-150。

Strapparava和r . Mihalcea(2008年)。学习识别文本中的情感，2008年ACM应用计算研讨会会议录(第1556-1560页)。纽约州纽约市:美国计算机学会。

Strapparava和a . Valitutti(2004年)。wordnet-affect:wordnet的情感扩展。《第四届语言资源与评估国际会议论文集》(第1083-1086页)。里斯本。

唐，洪，谭，程，谢(2009)。评论情感检测综述。专家系统与应用:国际期刊，36(7)，10760-10773。

Thelwall，M. (2009年)。MySpace评论。在线信息评论，33(1)，58-76。

塞尔沃尔博士、威尔金森博士和乌帕尔博士(2010年)。社交网络交流中的情感数据挖掘:MySpace中的性别差异。美国信息科学与技术学会杂志，21(1)，190-199。

瑟洛，C. (2003)。一代Txt？年轻人短信的社会语言学。话语分析在线，1(1)，检索于2008年1月3日，网址:http://extra . Shu . AC . uk/daol/articles/v 2001/n 2001/a 2003/thurlow 2002003-paper . html。

特尼博士(2002年)。竖起大拇指还是竖起大拇指？语义导向在无监督评论分类中的应用。《计算语言学协会(ACL)第40届年会论文集》，2002年7月6-12日，宾夕法尼亚州费城，417-424页。

j .瓦尔特和m .帕克斯(2002年)。过滤掉的线索，过滤掉的线索:以计算机为媒介的交流和关系。在m .纳普，j .戴利和g .米勒(编辑。)，人际交往手册(第3版。)(第529—563页)。千橡，加州:鼠尾草。

沃森博士(1988年)。积极和消极影响的个体内和个体间分析:它们与健康抱怨、感知压力和日常活动的关系。人格与社会心理学杂志，54卷6期，1020-1030页。

沃森博士，克拉克，洛杉矶，(1988年)。积极和消极情感简易测量方法的发展和验证:PANAS量表。人格与社会心理学杂志，54卷6期，1063-1070页。

威伯、威尔逊、布鲁斯、贝尔和马丁(2004年)。学习主观语言。计算语言学，30(3)，277-308。

韦贝、t .威尔逊和c .卡迪尔(2005年)。用语言注释观点和情感的表达。语言资源与评价，39(2-3)，165-210。

t .威尔逊(2008年)。精细的主观性和情感分析:识别私有状态的强度、极性和态度。匹兹堡大学。

茨韦塔纳·威尔逊、韦贝、j .、&霍夫曼出版公司(2009年)。识别上下文极性:短语级情感分析特征的探索。计算语言学，35(3)，399-433页。

Wilson，Wiebe，j .，& Hwa，R. (2006年)。识别强势和弱势观点从句。计算智能，22(2)，73-99。

威滕和弗兰克(2005年)。实用机器学习工具和技术。旧金山:摩根·考夫曼。

吴春辉，庄志坚，林永春(2006)。使用语义标签和可分离混合模型的文本情感识别。ACM亚洲语言信息处理汇刊，5(2)，165-183。

1. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558. Copyright © 2010 (American Society for Information Science and Technology) [↑](#footnote-ref-2)