Mood Patterns and Affective Lexicon Access in Weblogs

Thin Nguyen
Curtin University of Technology
Bentley, WA 6102, Australia
thin.nguyen@postgrad.curtin.edu.au

Abstract

The emergence of social media brings chances, but also challenges, to linguistic analysis. In this paper we investigate a novel problem of discovering patterns based on emotion and the association of moods and affective lexicon usage in blogosphere, a representative for social media. We propose the use of normative emotional scores for English words in combination with a psychological model of emotion measurement and a nonparametric clustering process for inferring meaningful emotion patterns automatically from data. Our results on a dataset consisting of more than 17 million mood-groundtruthed blogposts have shown interesting evidence of the emotion patterns automatically discovered that match well with the core-affect emotion model theorized by psychologists. We then present a method based on information theory to discover the association of moods and affective lexicon usage in the new media.

1 Introduction

Social media provides communication and interaction channels where users can freely participate in, express their opinions, make their own content, and interact with other users. Users in this new media are more comfortable in expressing their feelings, opinions, and ideas. Thus, the resulting user-generated content tends to be more subjective than other written genres, and thus, is more appealing to be investigated in terms of subjectivity and sentiment analysis. Research in sentiment analysis has recently attracted much attention (Pang and Lee, 2008), but modeling emotion patterns and studying the affective lexicon used in social media have received little attention.

Work in sentiment analysis in social media is often limited to finding the sentiment sign in the dipole pattern (negative/positive) for given text. Extensions to this task include the three-class classification (adding neutral to the polarity) and locating the value of emotion the text carries across a spectrum of valence scores. On the other hand, it is well appreciated by psychologists that sentiment has much richer structures than the aforementioned simplified polarity. For example, emotion – a form of expressive sentiment – was suggested by psychologists to be measured in terms of *valence* and *arousal* (Russell, 2009). Thus, we are motivated to analyze the sentiment in blogosphere in a more fine-grained fashion. In this paper we study the grouping behaviors of the emotion, or emotion patterns, expressed in the blogposts. We are inspired to get insights into the question of whether these structures can be discovered directly from data without the cost of involving human participants as in traditional psychological studies. Next, we aim to study the relationship between the data-driven emotion structures discovered and those proposed by psychologists.

Work on the analysis of effects of sentiment on lexical access is great in a psychology perspective. However, to our knowledge, limited work exists to examine the same tasks in social media context.

The contribution in this paper is twofold. To our understanding, we study a novel problem of emotion-based pattern discovery in blogosphere. We provide an initial solution for the matter using a combination of psychological models, affective norm scores for English words, a novel feature representation scheme, and a nonparametric clustering to automatically group moods into meaningful emotion patterns. We believe that we are the first to consider the matter of data-driven emotion pattern discovery at the scale presented in this

Proceedings of the ACL 2010 Student Research Workshop, pages 43–48,
Uppsala, Sweden, 13 July 2010.

paper. Secondly, we explore a novel problem of detecting the mood – affective lexicon usage correlation in the new media, and propose a novel use of a term-goodness criterion to discover this sentiment – linguistic association.

2 Related Work

Much work in sentiment analysis measures the value of emotion the text convey in a continuum range of valence (Pang and Lee, 2008). Emotion patterns have often been used in sentiment analysis limited to this one-dimensional formulation. On the other hand, in psychology, emotions have often been represented in dimensional and discrete perspectives. In the former, emotion states are conceptualized as combinations of some factors like valence and arousal. In contrast, the latter style argues that each emotion has a unique coincidence of experience, psychology and behavior (Mauss and Robinson, 2009). Our work utilizes the dimensional representation, and in particular, the core-affect model (Russell, 2009), which encodes emotion states along the valence and arousal dimensions. The sentiment scoring for emotion bearing words is available in a lexicon known as Affective Norms for English Words (ANEW) (Bradley and Lang, 1999). Related work making use of ANEW includes (Dodds and Danforth, 2009) for estimating happiness levels in three types of data: song lyrics, blogs, and the State of the Union addresses.

From a psychological perspective, for estimating mood effects in lexicon decisions, (Chastain et al., 1995) investigates the influence of moods on the access of affective words. For learning affect in blogosphere, (Leshed and Kaye, 2006) utilizes Support Vector Machines (SVM) to predict moods for coming blog posts and detect mood synonymy.

3 Moods and Affective Lexicon Access

3.1 Mood Pattern Detection

Livejournal provides a comprehensive set of 132 moods for users to tag their moods when blogging. The provided moods range diversely in the emotion spectrum but typically are observed to fall into soft clusters such as happiness (*cheerful* or *grateful*) or sadness *(discontent* or *uncomfortable)*. We call each cluster of these moods an *emotion pattern* and aim to detect them in this paper.

We observe that the blogposts tagged with moods in the same emotion pattern have similar

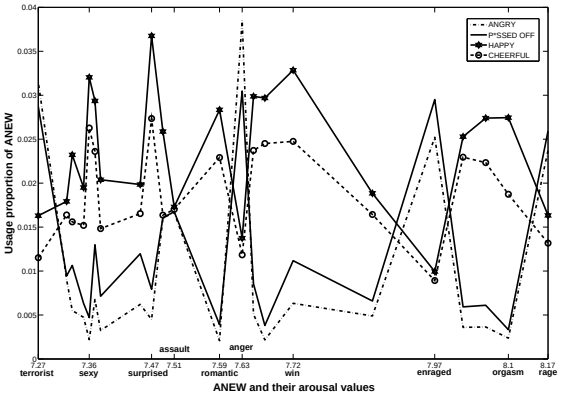

Figure 1: ANEW usage proportion in the posts tagged with *happy/cheerful* and *angry/p*ssed off*

proportions in the usage of ANEW. For example, in Figure 1 – a plot of the usage of ANEW having arousal in the range of 7.2 – 8.2 in the blogposts – we could see that the ANEW usage patterns of *happy/cheerful* and *angry/p*ssed off* are well separated. *Anger*, *enraged*, and *rage* will be most likely found in the *angry/p*ssed off* tagged posts and least likely found in the *happy/cheerful* ones. In contrast, the ANEW as *romantic* or *surprised* are not commonly used in the posts tagged with *angry/p*ssed off* but most popularly used in the *happy/cheerful* ones; suggesting that, the similarity between ANEW usage patterns can be used as a basis to study the structure of mood space.

Let us denote by $\mathcal{B}$ the corpus of all blogposts and by $\mathcal{M}$= {*sad, happy*, ...} the predefined set of moods ($|\mathcal{M}| = 132$). Each blogpost $b \in \mathcal{B}$ in the corpus is labeled with a mood $l_b \in \mathcal{M}$. Denote by n the number of ANEW ($n = 1034$). Let $\boldsymbol{x}^m = [\boldsymbol{x}_1^m, ..., \boldsymbol{x}_i^m, ..., \boldsymbol{x}_n^m]$ be the vector representing the usage of ANEW by the mood m. Thus, $\boldsymbol{x}_i^m = \sum_{b \in \mathcal{B}, l_b = m} c_{ib}$, where c_{ib} is the counting of the ANEW i-th occurrence in the blogpost b tagged with the mood m. The usage vector is normalized so that $\sum_{i=1}^{n} \boldsymbol{x}_i^m = 1$ for all $m \in \mathcal{M}$. To discover the grouping of the moods based on the usage vectors we use a nonparametric clustering algorithm known as Affinity Propagation (AP) (Frey and Dueck, 2007). AP is desirable here because it automatically discovers the number of clusters as well as the cluster exemplars. The algorithm only requires the pairwise similarities between moods, which we compute based on the Euclidean distances for simplicity.

To map the emotion patterns detected to their psychological meaning, we proceed to measure

the sentiment scores of those $|\mathcal{M}|$ mood words. In particular, we use ANEW (Bradley and Lang, 1999), which is a set of 1034 sentiment conveying English words. The valence and arousal of moods are assigned by those of the same words in the ANEW lexicon. For those moods which are not in ANEW, their values are assigned by those of the nearest father words in the mood hierarchical tree[1], where those moods conveying the same meaning, to some extent, are in the same level of the tree. Thus, each member of the mood clusters can be placed onto the a 2D representation along the valence and arousal dimensions, making it feasible to compare with the *core-affect* model (Russell, 2009) theorized by psychologists.

3.2 Mood and ANEW Usage Association

To study the statistical strength of an ANEW word with respect to a particular mood, the information gain measure (Mitchell, 1997) is adopted. Given a collection of blog posts $\mathcal{B}$ consisting of those tagged or not tagged with a target class attribute mood *m*. The entropy of $\mathcal{B}$ relative to this binary classification is

$$\mathcal{H}(\mathcal{B}) = -p_{\oplus}\log_2(p_{\oplus}) - p_{\ominus}\log_2 p_{\ominus}$$

where $p_{\oplus}$ and $p_{\ominus}$ are the proportions of the posts tagged and not tagged with *m* respectively.

The entropy of $\mathcal{B}$ relative to the binary classification given a binary attribute *A* (e.g. if the word *A* present or not) observed is computed as

$$\mathcal{H}(\mathcal{B}|A) = \frac{|\mathcal{B}_{\oplus}|}{|\mathcal{B}|}\mathcal{H}(\mathcal{B}_{\oplus}) + \frac{|\mathcal{B}_{\ominus}|}{|\mathcal{B}|}\mathcal{H}(\mathcal{B}_{\ominus})$$

where $\mathcal{B}_{\oplus}$ is the subset of $\mathcal{B}$ for which attribute *A* is present in the corpus and $\mathcal{B}_{\ominus}$ is the subset of $\mathcal{B}$ for which attribute *A* is absent in the corpus.

The information gain of an attribute ANEW *A* in classifying the collection with respect to the target class attribute mood *m*, $IG(m, A)$, is the reduction in entropy caused by partitioning the examples according to the attribute *A*. Thus,

$$IG(m, A) = \mathcal{H}(\mathcal{B}) - \mathcal{H}(\mathcal{B}|A)$$

With respect to a given mood *m*, those ANEW having high information gain are considered likely to be associated with the mood. This measure, also often considered a term-goodness criterion, outperforms others in feature selection in text categorization (Yang and Pedersen, 1997).

[1] http://www.livejournal.com/moodlist.bml

4 Experimental Results

4.1 Mood Patterns

We use a large Livejournal blogpost dataset, which contains more than 17 million blogposts tagged with the predefined moods. These journals were posted from May 1, 2001 to April 23, 2005. The ANEW usage vectors of all moods are subjected to a clustering to learn emotion patterns. After running the Affinity Propagation algorithm, 16 patterns of moods are clustered as below (the moods in upper case are the exemplars).

1. CHEERFUL, ecstatic, jubilant, giddy, happy, excited, energetic, bouncy, chipper
2. PENSIVE, determined, contemplative, thoughtful
3. REJUVENATED, optimistic, relieved, refreshed, hopeful, peaceful
4. QUIXOTIC, surprised, enthralled, devious, geeky, creative, recumbent, artistic, impressed, amused, complacent, curious, weird
5. CRAZY, horny, giggly, high, flirty, hyper, drunk, naughty, dorky, ditzy, silly
6. MELLOW, pleased, satisfied, relaxed, content, anxious, good, full, calm, okay
7. GRATEFUL, loved, thankful, touched
8. AGGRAVATED, irritated, bitchy, annoyed, frustrated, cynical
9. ANGRY, p*ssed off, infuriated, irate, enraged
10. GLOOMY, jealous, envious, rejected, confused, worried, lonely, guilty, scared, pessimistic, discontent, distressed, indescribable, crushed, depressed, melancholy, numb, morose, sad, sympathetic
11. PRODUCTIVE, accomplished, working, nervous, busy, rushed
12. TIRED, sore, lazy, sleepy, awake, groggy, exhausted, lethargic, drained
13. NAUSEATED, sick
14. MOODY, disappointed, grumpy, cranky, stressed, uncomfortable, crappy
15. THIRSTY, nerdy, mischievous, hungry, dirty, hot, cold, bored, blah
16. EXANIMATE, intimidated, predatory, embarrassed, restless, nostalgic, indifferent, listless, apathetic, blank, shocked

Generally, the patterns 1–7 contain moods in high valence (pleasure) and the patterns 8–16 include mood in low valence (displeasure). To examine whether members in these emotion patterns

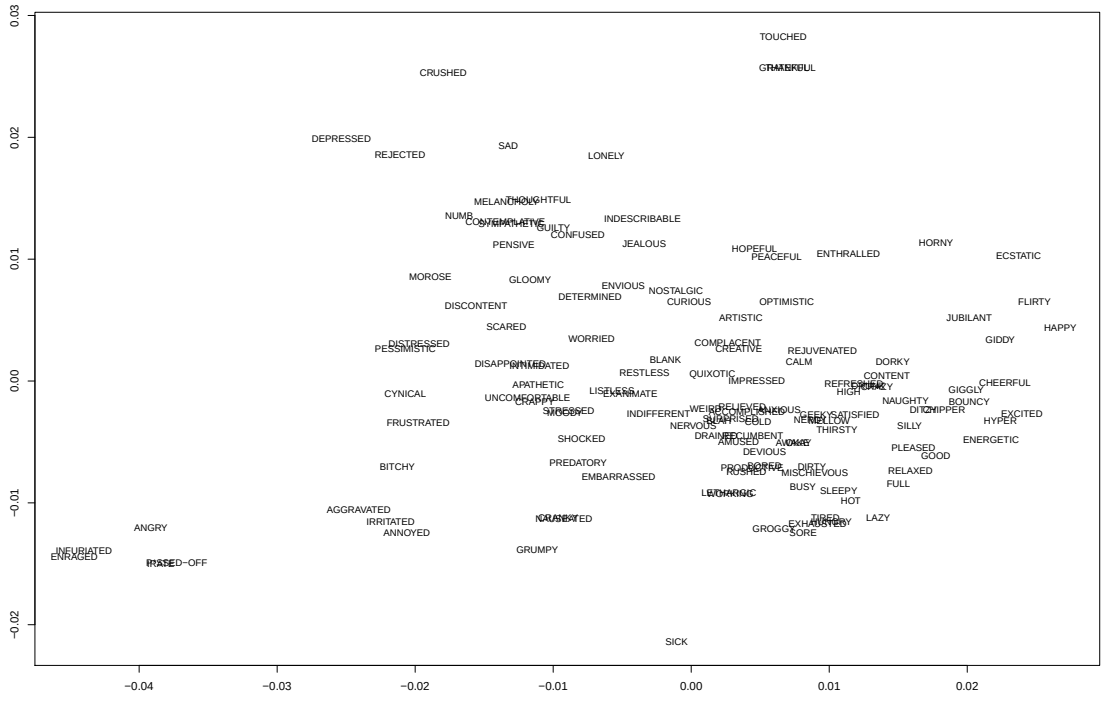

Figure 2: Projection of moods onto a 2D mesh using classical multidimensional scaling

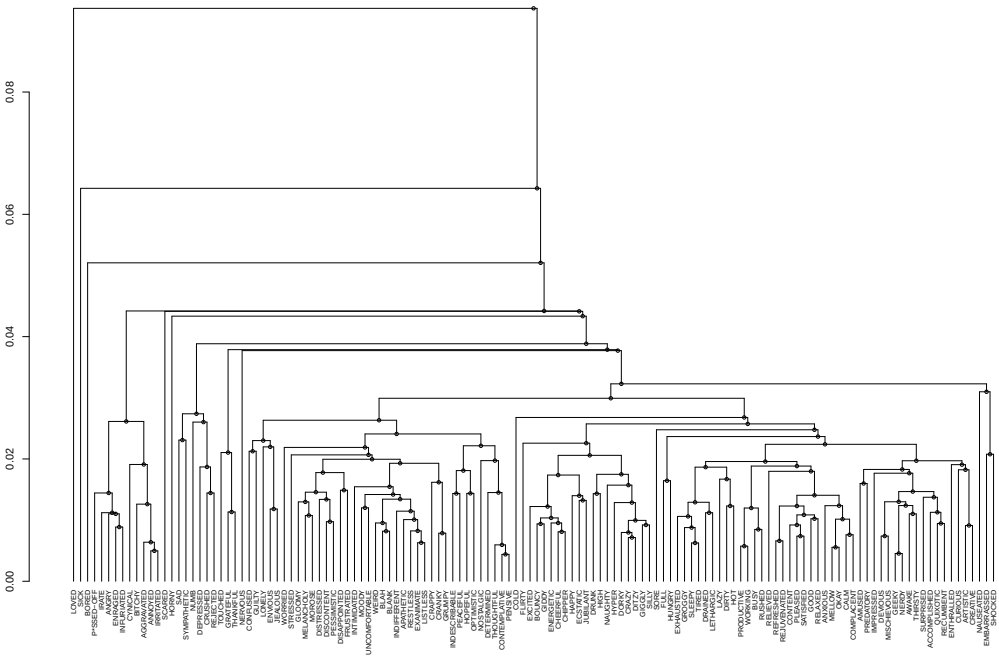

Figure 3: The clustered patterns in a dendrogram using hierarchical clustering

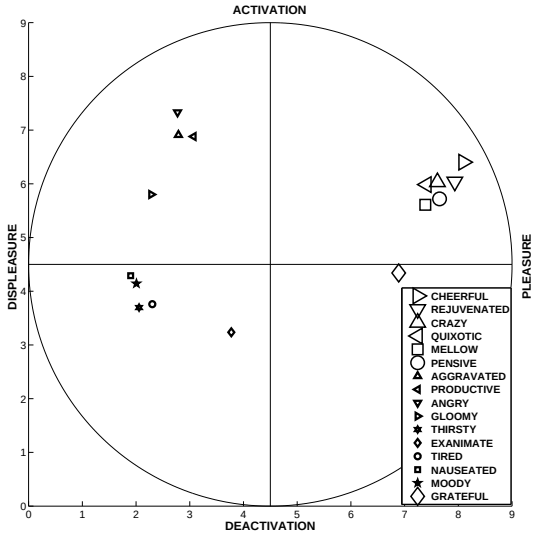

Figure 4: Discovered emotion patterns in the affect circle

follow an affect concept, we place them on the affect circle (Russell, 2009). We learn that nearly all members in the same patterns express a common affect concept. Those moods in the patterns with *cheerful*, *pensive*, and *rejuvenated* as the exemplars are mostly located in the first quarter of the affect circle (0^0 – 90^0), which should contain moods being high in both pleasure and activation measures. Meanwhile, many members of the *angry* and *aggravated* patterns are found in the second quarter (90^0 – 180^0), which roughly means that those moods express the feeling of sadness in the high of activation. The patterns with the exemplars *nauseated* and *tired* contain a majority of moods found in the third quarter (180^0 – 270^0), which could be representatives for the mood fashion of sadness and deactivation. In addition, the *grateful* group could be a representative for moods which are both low in pleasure and in the degree of activation (270^0 – 360^0 of the affect circle). Thus, the clustering process based on the ANEW usage could separate moods having similar affect scores into corresponding segments in the circle proposed in (Russell, 2009).

To visualize mood patterns that have been detected, we plot these emotion modes on the affect circle plane in Figure 4. For each pattern, the valence and arousal are computed by averaging of the values of those moods in the quarter where most of the members in the pattern are.

To further visualize the similarity of moods, the ANEW usage vectors are subject to a classical multidimensional scaling (Borg and Groenen, 2005) (MDS) and a hierarchical clustering. Figure 2 and Figure 3 show views of the distance between moods, based on the Euclidean measure of their corresponding ANEW usage, using MDS and hierarchical clustering respectively.

Mood	Top ANEW words associated
Cheerful	fun, happy, hate, good, christmas, merry, birthday, cute, sick, love
Happy	happy, hate, fun, good, birthday, sick, love, mind, alone, bored
Angry	angry, hate, fun, mad, love, anger, good, stupid, pretty, movie
P*ssed off	hate, stupid, mad, love, hell, fun, good, god, pretty, movie
Gloomy	sad, depressed, hate, wish, life, alone, lonely, upset, pain, heart
Sad	sad, fun, heart, upset, wish, funeral, hurt, pretty, loved, cancer

(a) Moods and the most associated ANEW words

ANEW	Most likely moods	Least likely moods
Desire	contemplative, thoughtful	enraged, drained
Anger	angry, p*ssed off	nauseated, grateful
Accident	sore, bored	exanimate, indifferent
Terrorist	angry, cynical	rejuvenated, touched
Wine	drunk, p*ssed off	ditzy, okay

(b) ANEW words and the most associated moods

Table 1: Mood and ANEW correlation

4.2 Mood and ANEW Association

Based on the IG values between moods and ANEW, we learn the correlation of moods and the affective lexicon. With respect to a given mood, those ANEW having high information gain are most likely to be found in the blogposts tagged with the mood. The ANEW most likely happened in the blogposts tagged with a given mood are shown in Table 1a; the most likely moods for the blog posts containing a given ANEW are shown in Table 1b.

The ANEW used in the blog posts tagged with moods in the same pattern are more similar than those in the posts tagged with moods in different patterns. In Table 1a, the most associated ANEW

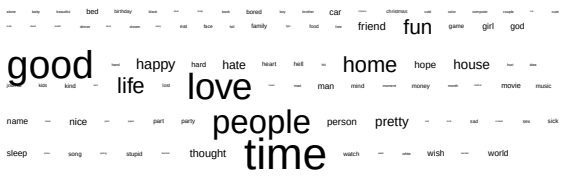

Figure 5: Top 100 ANEW words used in the dataset

in the blogposts tagged with *cheerful* are more similar to those in *happy* ones than those in *angry* or *p*ssed off* ones.

For a given mood, a majority of the ANEW used in the blog posts tagged with the mood is similar in the valence with the mood. The occurrence of some ANEW having valence much different with the tagging mood, e.g. the ANEW *hate* in the posts tagged with *cheerful* or *happy* moods, might be the result of a negation construction used in the text or of other context.

For a given ANEW, the most likely moods tagged to the blog posts containing the word are similar with the word in the affective scores. In addition, the least likely moods are much different with the ANEW in the affect measure. A plot of top ANEWs used in the blogposts is shown in Figure 5.

Other than the ANEW conveying abstract concept, e.g. *desire* or *anger*, those ANEW expressing more concrete existence, e.g. *terrorist* or *accident*, might be a good source for learning opinions from social network towards the things. In the corpus, the posts containing the ANEW *terrorist* are most likely tagged with *angry* or *cynical* moods. Also, the posts containing the ANEW *accident* are most likely tagged with *bored* and *sore* moods.

5 Conclusion and Future Work

We have investigated the problems of emotion-based pattern discovery and mood – affective lexicon usage correlation detection in blogosphere. We presented a method for feature representation based on the affective norms of English scores usage. We then presented an unsupervised approach using Affinity Propagation, a nonparametric clustering algorithm that does not require the number of clusters a priori, for detecting emotion patterns in blogosphere. The results are showing that those automatically discovered patterns match well with the core-affect model for emotion, which is independently formulated in the psychology literature. In addition, we proposed a novel use of a term-goodness criterion to discover mood–lexicon correlation in blogosphere, giving hints on predicting moods based on the affective lexicon usage and vice versa in the social media. Our results could also have potential uses in sentiment-aware social media applications.

Future work will take into account the temporal dimension to trace changes in mood patterns over time in blogosphere. Another direction is to integrate negation information to learn more cohesive association in affect scores between moods and affective words. In addition, a new affective lexicon could be automatically detected based on learning correlation of the blog text and the moods tagged.

References

I. Borg and P.J.F. Groenen. 2005. *Modern multidimensional scaling: Theory and applications*. Springer Verlag.

M.M. Bradley and P.J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical report, University of Florida.

G. Chastain, P.S. Seibert, and F.R. Ferraro. 1995. Mood and lexical access of positive, negative, and neutral words. *Journal of General Psychology*, 122(2):137–157.

P.S. Dodds and C.M. Danforth. 2009. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, pages 1–16.

B.J. Frey and D. Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814):972.

G. Leshed and J.J. Kaye. 2006. Understanding how bloggers feel: recognizing affect in blog posts. In *Proc. of ACM Conf. on Human Factors in Computing Systems (CHI)*.

I.B. Mauss and M.D. Robinson. 2009. Measures of emotion: A review. *Cognition & emotion*, 23:2(2):209–237.

T. Mitchell. 1997. *Machine Learning*. McGraw Hill.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

J.A. Russell. 2009. Emotion, core affect, and psychological construction. *Cognition & Emotion*, 23:7(1):1259–1283.

Y. Yang and J.O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proc. of Intl. Conf. on Machine Learning (ICML)*, pages 412–420.