T2PP Workshop, 9-10 April 2010, Vrije Universiteit Amsterdam

Leveraging Textual Sentiment Analysis with Social Network Modelling:

Sentiment Analysis of Political Blogs in the 2008 U.S. Presidential Election

Wojciech Gryc
Oxford Internet Institute, University of Oxford
1 St Giles
Oxford OX1 3JS
United Kingdom
wojciech.gryc@oii.ox.ac.uk

and

Karo Moilanen
Oxford University Computing Laboratory
Wolfson Building, Parks Road
Oxford OX1 3QD
United Kingdom
karo.moilanen@comlab.ox.ac.uk

Abstract

Automatic computational analysis and categorisation of political texts with respect to the rich array of personal sentiments, opinions, stances, and political orientations expressed in polarised political discourse is an exciting task which opens up many avenues for more accurate and naturalistic large-scale political analysis. The task does however pose major challenges for state-of-the-art Sentiment/Subjectivity/Affect Analysis and general Natural Language Processing tools. In this initial study, we investigate the feasibility of combining purely linguistic indicators of political sentiment with non-linguistic evidence gained from concomitant social network analysis. This study focuses on political blog analysis and draws on a corpus of 2.8 million blog posts by 16,741 bloggers crawled between April 2008 and May 2009. We focus on modelling blogosphere sentiment centered around Barack Obama during the 2008 U.S. presidential election, and describe a series of initial sentiment classification experiments on a data set of 700 crowd-sourced posts labelled as ‘positive’, ‘negative’, ‘neutral’, or ‘not applicable’ with respect to Obama. Our approach employs a hybrid machine learning and logic-based framework which operates along three distinct levels of analysis encompassing standard shallow document classification, deep linguistic multi-entity sentiment analysis and scoring, and social network modelling. The initial results highlight the inherent complexity of the classification task, and point towards the positive effects of learning features that exploit entity-level sentiment and social network structure.

Keywords: political blogs, social networks, entity-level and document-level sentiment analysis

1 Introduction

Political blogs constitute a fascinating genre. They are typically charged with extremely high amounts of personal opinions, sentiments, stances, feelings, and emotions towards and about a multitude of individuals, organisations, issues, and events that have some political relevance, and represent a very large number of people. Due to the sheer size of such a complex distributed, dynamic, and vibrant information space, the only viable way of monitoring, analysing, and predicting information in the political blogosphere is to develop a large-scale computational multi-paradigm and multi-modal framework that can deal with the content as well as the social aspects of the blogosphere.

With this lofty goal in mind, we investigate in this study the feasibility of combining purely textual indicators of political sentiment with non-textual information gained from concomitant social network analysis. We focus on political blog analysis and base our research on a large corpus of posts by 16,741 bloggers crawled daily between April 2008 and May 2009. Rich political blog data allows us to explore a number of major themes in political sentiment analysis, social network analysis, text mining, and, most importantly, how these areas interact. For example, by exploring the content of each post and how it fits into the larger social network in which it resides, one can see that discussions tend to cluster differently depending on individual politicians: people discussing Barack Obama tend to be more dispersed than those discussing John McCain, for example. An interesting and highly relevant research question is whether such network-based information can be used to improve the accuracy of automatic political sentiment classification.

In this initial study, we approach automatic sentiment analysis of political blogs using a hybrid machine learning and logic-based classification framework which operates along three distinct complementary levels of analysis, each capable of offering a unique representation of each blog post:

- **Shallow Document Classification** (§4.1): as a first, strong baseline centered around holistic document-level textual evidence alone, each blog post is represented using standard *n*-gram features.
- **Deep Entity-level Sentiment Scoring** (§4.2): in a second approach centered around much more focused lower-level sentiment evidence, each blog post is represented as a set of detailed sentiment scores assigned to all individual entities (e.g. mentions of politicians, places, organisations, and abstract issues) mentioned in the post.
- **Social Network Modelling** (§4.3): in a third, much wider meta-approach centered around the blog posts' social context, each blog post's position in the social network structure across the whole blog post space is used as classification evidence.

Classification Task. Although our framework is not restricted to any particular topic or entity (be they concrete or abstract), we confine[1] ourselves in this initial study to studying only one specific entity in order to gain a better understanding of the problem. In light of the importance of social media to Barack Obama's political campaign in the 2008 U.S. presidential election, we decided to focus our analysis on the sentiment expressed in the blogosphere towards and about Barack Obama within the time period that our blog data represents. The classification task that we attempt in this initial study can be summarised specifically as follows:

- *Given a political blog post, does it, as a whole, express positive, neutral, or negative sentiment towards or about the target entity (Barack Obama)?*

The above classification task can accordingly be characterised as **entity-centric document-level sentiment classification** in that the document-level sentiment polarity label of a given post reflects the overall sentiment towards or about a single target entity in that post, not the overall sentiment of the post as a whole. Note that this type of sentiment classification is different from the 'traditional' document-level classification paradigm because a post that is negative *overall* (i.e. as a document) can (and is likely to) be concurrently negative, positive, or neutral towards or about many individual entities, for example. We approach the problem as if we were trying to label the entire data set after it was collected, rather than in real time. We are more interested in being able to label a majority of the 2.8 million posts we have with some level of confidence, rather than having to look into the future when labeling specific posts.

General Challenges. Automatic computational analysis and categorisation of political texts is on the whole a seriously challenging task, as has been observed in the area (e.g. [4], [15], [19], [23], [25]). When the analytical scope is extended to include further non-factual aspects of meaning pertaining to subjectivity, sentiment, opinions, affect, and emotions, the analytical and computational challenges become even more pronounced. This is especially true with politically charged content in blogs because, as a genre, political blogs represent noisy, in-depth, collaborative, and dynamic discussions and debates by multiple contributors across a wealth of topics, issues, and entities - only some of which constitute core content while some others are mere digressing or tangential content. Blog posts further link to each other via highly complex interrelated direct/explicit and indirect/implicit structural, semantic, rhetorical, and temporal chains. It can be argued that no such chain can be explained fully out of context, although blog posts are likely to carry at least some clues that an algorithm can exploit. In addition to their distributed nature, blog posts can

[1] We aim to expand this in the future to further politicians and topics such as *"McCain"* and *"Iraq"*.

also include other forms of multimedia such as embedded videos or images which existing NLP algorithms cannot easily align with the text content.

From the viewpoint of textual sentiment analysis algorithms, any classification evidence that may be gleaned from blog posts is inherently noisy as bloggers' real sentiments and opinions are often obfuscated by complex rhetoric, irony, sarcasm, comparisons, speculation, and other paralinguistic devices that prototypically characterise political discussions. In addition, domain-specific terms, word senses, and vocabularies, and informal/non-standard registers also feature frequently. As is the case with Web content in general, political blogs also come with many purely textual hurdles that have to be faced by NLP tools such as complex or incomplete grammatical structures, broken sentence boundaries, quotes, junk characters, and spelling anomalies.

Even if such structural and textual problems were solved fully, the very task of automatically detecting bloggers' political sentiments, opinions, and orientation pertaining to highly polarised political issues would still remain formidable. This stems from the fact that current computational tools - be they linguistic or non-linguistic - struggle to map raw surface clues onto deeper semantic representations which ultimately require, *for each blogger and for each issue, bill, or event* under consideration, the ability (i) to detect the blogger's political party affiliations, political viewpoints, professional background, motivation, general knowledge, and, indeed, affective states; (ii) to measure the blogger's political extremeness or distance from a centrist position; (iii) to measure the blogger's confidence and agreeability/argumentativeness in the discussions; (iv) to measure how important something is to the blogger politically; (v) to understand how meaning was constructed collaboratively by the bloggers; (vi) to understand why certain topics are (not) discussed; and (vii) to detect sincere opinions vs. deliberate flaming. Moreover, not only are many political opinions latent behind expression which (from an algorithm's point of view) present themselves as purely neutral but some explicit political opinion expressions may even be inversely related to the blogger's actual political orientation (after *ibid.*).

2 Data

2.1 Election Data

The present initial study, which is part of a wider research project, focuses on data provided by IBM's *Predictive Modeling Group*. In total, the data set consists of 2,782,356 posts written by 16,741 politically-oriented bloggers collected between April 22 2008 and May 1 2009 (many of which focus on the U.S. Presidential election in 2008). The included blogs were chosen based on the tags were associated with them on the *Technorati*[2] blog indexing service. Such a blog data set provides an interesting opportunity for researchers because it contains text-based discussions on politics as well as date stamps and hyperlinks between blog posts. The hyperlink feature can be treated as a citation network which shows the various ways in which individual bloggers are and become aware of each other, and how information flows within the political blogosphere.

Posts were found through blogs self-reporting new content through the RSS (*Real Simple Syndication*) standard which includes hyperlinks to the content posted on the blogs themselves. Once the blogs were crawled, the content of each post was filtered to discard text and hyperlinks from advertisements, side bar content (e.g. blog rolls), and other portions of the web pages that include unwanted superfluous content.

The role of content filtering is extremely important, mainly for three reasons. Firstly, blog titles and content from side bars can easily bias any statistical models trained on noisy data. Since we are only interested in the main content of every blog post, it is important to be able to avoid such bias by discarding as many sidebar links as possible. Secondly, superfluous content can, even in structural terms, cause potentially devastating complications for NLP tools which can manifest themselves as incorrect sentence breaking, part-of-speech tags, phrase chunking, and parsing - all of which will deteriorate the performance of sentiment classifiers that use NLP components. Lastly, blogs often acted in an automated fashion. For example, blog A citing blog B can automatically cause blog A to leave a link as a comment on blog B's post and hence result in additional links. Blog rolls, while providing useful social network information, may cause the network to become

[2] http://technorati.com/

artificially dense with hyperlinks that do not pertain to the blog posts themselves. It is important to note, however, that while our filtering strategy is effective enough for practical purposes, it is based on heuristics and is by no means perfect.

2.2 Sentiment Annotations

While the aforementioned data collection process allows us to analyse the content of the posts and to build and model social networks of bloggers and their posts, the crawling process did not as such deal with the sentiment expressed in the posts. In order to gain access to the sentiment properties of the posts, we instead made use of Amazon's *Mechanical Turk*[3] service which involves *Requesters* (us) posting batches of small tasks known as HITs (*Human Intelligence Tasks*) which are completed by *Workers* (*Turkers*) for a small monetary reward ($0.03 per blog post).

For the present initial study, we selected a random sample of 700 blog posts discussing Barack Obama and asked the Turkers to label them as (i) POSITIVE, NEGATIVE, NEUTRAL towards or about Obama, or (ii) NOT APPLICABLE (with respect to Obama). In order for us to monitor and ensure the post-level consistency of the Turkers' sentiment ratings, we required each post to be labelled by three Turkers. In total, 86 unique Turkers took part in the task. All posts labelled as NOT APPLICABLE were discarded because such cases typically indicate that a given blog post to be labelled had been taken offline or contained irrelevant non-textual content (e.g. a video or an image). From the resultant raw sentiment ratings, two labelled subsets were generated as follows:

- **Lenient Majority Vote**: The first subset contains all posts that received a majority vote whereby either 2/3 or 3/3 Turkers had to agree on the label of each post. This resulted in 454 posts from which a final[4] labelled corpus of 439 posts was obtained. All results reported in the present initial study are from this subset.

- **Strict Agreement**: The second subset contains all posts that received unanimous 3/3 votes. Since that criterion resulted in only 124 posts, this subset is not included in the present study.

We are in the process of increasing the amount of sentiment annotations.

Human Performance. Even though it is complicated to estimate the inter-annotator agreement rates between 86 annotators (each of whom provided different amounts of annotations), some tentative observations can be made regarding the expected human agreement and performance ceiling in the sentiment classification task that our classifiers aim at solving. Taking all ratings on which 2/3 or 3/3 of the annotators agreed (excluding all NOT APPLICABLE cases), Table 1 shows the distribution of votes per sentiment polarity. It interestingly reveals that only 126 (27.75%) display FULL AGREEMENT, 145 (31.94%) display disagreements that involve neutral polarity (NEUTRAL DISAGREEMENT), and 183 (40.31%) display FATAL DISAGREEMENT in the form of opposing non-neutral polarities. Both the noticeably low amount of FULL AGREEMENT and the relatively high amount of FATAL DISAGREEMENT suggest that, perhaps not surprisingly, the 3-way classification task is highly subjective even for humans. It is against the strict ceiling of 27.75% (or the lenient one without FATAL DISAGREEMENT cases at 59.69%) that the classifiers' performance ought to be compared.

Data Quality. This kind of empirical data crowd-sourcing offers practical benefits that cannot be refuted. It can further be argued that, because they reflect real, uncontrolled, and 'untrained' opinions, crowd-sourced sentiment annotations are maximally valid in terms of their naturalness. The downside is naturally that the quality of the resultant annotations may be lower than what can be reached in traditional, more rigorous controlled and vetted annotation campaigns. In particular, the very nature of the MT service is based on the notion of quantity rather than quality as the Turkers expect simple tasks that do not require any comprehensive annotator training as such[5] so that each can be completed in a matter of seconds to reflect the typically paltry per-item pay rates.

Despite these quality concerns, the use of crowd-sourcing as a data collection method has proven effective for machine learning in general and sentiment analysis in particular. Hsueh et al.

[3] `https://www.mturk.com/mturk/welcome`

[4] A small amount of posts were excluded due to anomalous content and features.

[5] A typical HIT page contains only simple instructions and/or examples.

Table 1: Distribution of 3-way sentiment judgements from 86 annotators

POS	NTR	NEG	#	%	Agreement Type	#	%
	NTR (3)		46	10.13%	FULL AGREEMENT		
		NEG (3)	45	9.91%	FULL AGREEMENT		
POS (3)			34	7.49%	FULL AGREEMENT		
	NTR (2)		1	0.22%	FULL AGREEMENT		
						126	27.75%
POS (1)	NTR (2)		43	9.47%	NEUTRAL DISAGREEMENT		
	NTR (2)	NEG (1)	40	8.81%	NEUTRAL DISAGREEMENT		
	NTR (1)	NEG (2)	39	8.59%	NEUTRAL DISAGREEMENT		
POS (2)	NTR (1)		23	5.07%	NEUTRAL DISAGREEMENT		
						145	31.94%
POS (1)		NEG (2)	82	18.06%	FATAL DISAGREEMENT		
POS (2)		NEG (1)	56	12.33%	FATAL DISAGREEMENT		
POS (1)	NTR (1)	NEG (1)	44	9.69%	FATAL DISAGREEMENT		
POS (1)		NEG (1)	1	0.22%	FATAL DISAGREEMENT		
						183	40.31%
			454	100.00%			

[8], for example, carried out an analysis of sample post snippets from the same pool of blog data as ours and confirmed the high quality of ratings from Turkers against those from 'expert' annotators. Turker annotations have also been used to rate Wikipedia articles [10] and news headlines [22]: in both studies, the quality[6] of Turker ratings was also comparable to that of expert annotators.

We hence conclude that the seemingly low inter-annotator agreement rates on our data set were not the byproduct of crowd-sourced annotations as such but rather reflect the inherently fuzzy and subjective properties of the underlying sentiment classification task.

3 Related Work

Due to the fact that our classification framework involves tools, ideas, and phenomena from multiple paradigms, topics, areas, and fields, space limits prevent us from surveying all of them. We hence limit the discussion on the relation of the present study with past work and existing proposals to **political sentiment analysis**.

The recent surge of interest in 'mainstream' (e.g. product/movie review-oriented) Sentiment Analysis and Opinion Mining[7] has touched upon the political domain in the form of a few studies which have followed the standard document topic classification paradigm and which have simply mapped the default positive-neutral-negative sentiment polarities onto political 'polarities' in bi- or tripartite political systems. These document-level approaches typically use some form of machine learning with no or only shallow linguistic features. [19] discuss the application of sentiment analysis to informal political discourse to predict political affiliations as RIGHT (Republican, conservative, r-fringe) vs. LEFT (Democrat, liberal, l-fringe) in blog posts using a probabilistic classifier (accuracy $\sim$60.37%). In a similar study, [15] used web-based Pointwise Mutual Information scoring, supervised machine learning, and citation graph clustering (accuracy 68.48%$\sim$73%).

Other variants of the same paradigm have focused on classifying public comments on proposed governmental regulations as PRO vs. AGAINST with a combination of sentiment analysis, (sub)topic detection, argument structure analysis, and semantic frame analysis ([11]). The cultural orientation and ideologies in left- and right-wing political documents were estimated based on co-citation information in [5] (accuracy $\sim$90%) while congressional floor debates were classified as SUPPORT

[6] Anecdotal evidence and general opinions amongst the users of crowd-sourced annotations however suggest that their quality depends (sometimes entirely randomly) on the annotation task attempted and the pool of Turkers that participated in it, and that the risk of obtaining junk annotations is always present.

[7] For a recent survey of these areas, see [20].

FOR vs. OPPOSITION TO to (a piece of legislation) using graph-based agreement links between speech segments and speaker identities in [23] (accuracy ∼70.81%). Others have involved various forms of supervised learning to classify congressional floor debates for general sentiment ([25]) (accuracy ∼65.5%); to capture what kinds of subjective perspectives (points of view) are expressed in text pertaining to the ISRAELI vs. PALESTINIAN polar classes ([14]) (accuracy 93.46∼99.09% for documents, ∼94.93% for sentences); and to classify LEFT-VOICE vs. RIGHT-VOICE blog posts about President Bush's management of the Iraq War ([4]) (accuracy ∼89.77%), amongst others.

Further similar approaches can be found in the form of predictive models which include temporal features. An opinion forecasting approach was described in [12] who combine a shallow bag-of-words approach with predefined entities, syntactic parsing, and temporal news coverage models to predict the impact of news on public perception of political candidates. [9] in turn describe a supervised learning system that predicts which party is going to win the election on the basis of opinions posted on an election prediction website (accuracy ∼81.68%).

While these studies have reported relatively high 'accuracy' levels, the standard document topic classification paradigm is too coarse to score individual entities. To the best of our knowledge, the role of deeper sentiment analysis in general and fine-grained multi-entity scores in particular has not been investigated fully in the area of political sentiment analysis.

4 Overview of the Classification Framework

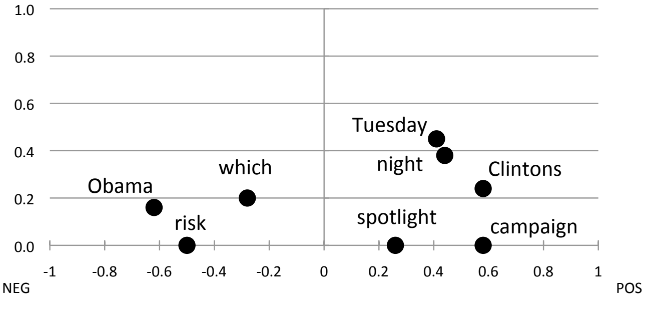
4.1 Shallow Document Classification

One of the simplest approaches to text categorisation is arguably the bag-of-words paradigm in which a given text is represented as an unordered collection of independent statistical features pertaining to word or *n*-gram frequencies. Although structural and positional information about words and *n*-grams is discarded altogether, the bag-of-words method is hard to beat in practice. We therefore adopt as a strong baseline classifier a unigram count model ([16]) that operates on stemmed[8] and normalised *n*-grams from the blog posts. Rather than using raw word frequencies directly, we use TF-IDF vectors to represent the posts themselves. We limited the word vectors to terms that appear in at least 5 different posts, and avoided terms that appeared in over 500 posts. This resulted in 4647 unigram features for classification.

4.2 Deep Entity-level Sentiment Scoring

In order to complement with deeper linguistic and sentiment information the holistic evidence offered by the shallow *n*-gram method, we employed a wide-coverage sentiment parser. In particular, we wished to utilise in the analysis information about the overall sentiment expressed in each blog post towards all individual entities mentioned in it. The parser, which is described in greater detail in [18], employs compositional sentiment logic, deep grammatical analysis, and large manually compiled sentiment lexica to exhaustively assign sentiment scores to different structural levels across individual words, syntactic phrases, sentences, and documents. In particular, it assigns gradient POS:NTR:NEG sentiment scores for all individual entity mentions (e.g. *“Obama$^{(+)}$”*, *“Obama's$^{(-)}$”*, *“Barack$^{(N)}$”*, *“Chicago$^{(+)}$”*, ...) and aggregated entity topics (e.g. *“Obama”* had 25 mentions 58% of which were positive) in a given blog post. Ex. 1 shows a sample sentence from our corpus and the gradient sentiment scores that the parser assigned to the [ENTITIES] in it:

(1) *Judging by* [TUESDAY] [NIGHT]*, the* [CLINTONS] *would want to share the* [CAMPAIGN] [SPOTLIGHT]*,* [WHICH] *runs the* [RISK] *of making Mr.* [OBAMA] *look weak.*



[8] As output by the standard Porter stemmer ([21]).

We transformed the aggregated topical POS:NTR:NEG sentiment percentage counts from all entities mentioned across all 700 posts into unique learning features (e.g. OBAMA_NEG_SCORE, PALIN_NTR_SCORE, IRAQ_NEG_SCORE, ...). Ex. 2 shows the top 25 most frequent topical entities from the posts.

(2) *Obama (115), that (97), you (91), he (91), it (91), barack (88), I (84), they (81), we (81), who (79), what (70), McCain (69), people (67), this (65), campaign (64), candidate (59), there (56), president (53), John (52), time (51), election (51), all (49), one (48), comment (48), day (46)*

In order to make the features more focused around the key entities and issues in the election, we discarded features from entities that had a frequency of <10 across our corpus. This filtering resulted in 951 entity score features for classification.

4.3 Social Network Modelling

In addition to serving as an interesting text corpus of political sentiment, our data set contains rich information about the relationships between individual bloggers represented by various hyperlinking structures. As many bloggers link to each other within their posts, such link data can reveal the rich underlying social structures in the political blogosphere. Past research into such linking patterns during the 2004 U.S. presidential election has, for example, shown that bloggers tend to segregate themselves on ideological grounds, with conservative and liberal bloggers separating into tight-knit clusters that have different behavioural characteristics with regards to hyperlinking to other blogs ([1]). Such segregation has also been observed in more topically focused blog communities such as war blogs ([24]). Analysis of links between ideologically-charged blog clusters similarly shows that links are often used to critique other bloggers ([7]).

Since blogs often self-categorise into ideologically-charged clusters, incorporating information about such clusters into our blog post categorisation model appears intuitively beneficial. We accordingly sought to investigate the possibility of leveraging the above kinds of social linking and sorting phenomena observed in the political blogosphere to facilitate the blog post classification and labelling task. In this initial study, we take a relatively simple approach to exploring which clusters bloggers find themselves in, with posts then acquiring the features of their parent blogs.

Weakly Connected Components. While one could look at the individual post-level linking patterns between blogs (e.g. [13]), the small number of post labels to which we currently have access - combined with sparse post-level networks - means that not enough post-level social information may be gleaned from the data. Social networks at the blog level do however provide more information as all hyperlinks observed between blogs in the data set can be aggregated into a directed network. We treated the connections between blogs as unweighted: in particular, if at any point during the year blog A was linked to blog B, then we treated this as a link in the blog graph. The one shortcoming of such an approach is that relationships between bloggers that regularly link to each other are treated the same as one-off links.

Using the *iGraph* package[9] ([3]), we determined for each blog its location in the social network alongside a number of different post-specific properties. In each case, we explored different subgraph types and whether posts were written by a blogger situated in specific subgraphs of the network. For example, a simple subgraph structure within a directed network is a *weakly connected component* which represents subgraphs where all nodes are connected to all other nodes. In our blog network, the largest such weakly-connected component consists of 8297 blogs, while the second-largest has 67. In this case, two variables could be added to the feature vector used to classify a given blog post, namely 1) is the parent blog situated in the largest component, and 2) is the parent blog situated in the second-largest component? From these, ten boolean features were generated that indicate which community a given post belongs to, based on which blog was responsible for posting it.

Community Detection Algorithms. A second plausible approach is afforded by community detection algorithms which aim at finding dense subgraphs within a social network. While a component can be one interpretation of a community within a social network, community-finding algo-

[9] http://igraph.sourceforge.net/

rithms often have more stringent definitions. The general idea behind many such algorithms is to locate parts of a social network that are more dense than one would expect based on the network as a whole. Such a dense subgraph may imply stronger relationships between members of the subgraph compared to members outside of the subgraph. If we assume that the social network is homophilous by ideology ([17]) - that is, if blogs tend to link to each other more often when they share similar political views - then we can use their membership within subgraphs as features for classification. Through the use of a *fast greedy community detection* algorithm ([2]), a number of communities were detected in the blog network. Fast greedy community detection places nodes into communities in a way that maximises the number of edges within communities, rather than between them. It is an agglomerative approach where each node begins by being within its own community, and communities are then merged to maximise within-community links and minimise between-community links. The ten largest communities found[10] using this approach are outlined in Table 2.

Table 2: The ranks and sizes of the largest communities found in the aggregated blog network.

Rank	1	2	3	4	5	6	7	8	9	10
Size	2077	1659	1131	945	582	484	380	355	98	82

From the above approaches to modelling the social network structure of our blog sample, 23 cluster features were generated for classification.

4.4 Overview of Algorithms

For all experiments, we used standard Naïve Bayes Multinomial (NBM) and Logistic Regression models available in the WEKA[11] toolkit ([6]), all with their default parameters. By assuming that individual terms appear in a blog post independently of all other terms, the NBM classifier calculates the probability of a given blog post belonging to a positive, neutral, or negative category. This is done by summing the estimated log-probabilities of individual terms appearing in the categories.

We further made use of 2nd-tier Logistic Regression metaclassifiers which use as their inputs the probabilistic predictions made by three 1st-tier NBM classifiers. As metaclassifiers, we compared two different options, namely 1) a Majority Voting classifier which counts the three class labels predicted by the 1st-tier classifiers and assigns the majority class label to a given blog post, and 2) a Stacking classifier which takes three inputs from the 1st-tier classifiers and treats them as features for the final classification step.

We further experimented with Support Vector Machines (SVMs) and J48 decision tree algorithms. Because they did not perform as well as the above classifiers on our data set, their results are not included in this paper.

5 Experiments

5.1 Experimental Conditions

We report the performance of three different feature types (§4) across 5620 features in the following conditions:

- (1) A Logistic Regression classifier with 23 social network features [SNA]
- (2) An NBM classifier with 951 sentiment analysis features [SA]
- (3) An NBM classifier with 4647 unigram bag-of-words features [BOW]
- (4) An NBM classifier with all 5620 features [ALL]

[10] Note that the blog network was symmetrised prior to running the fast greedy algorithm.

[11] `http://www.cs.waikato.ac.nz/ml/weka/`

- (5) A Stacking classifier with three separate NBM classifiers for (1), (2), and (3) [STACK]
- (6) A Voting classifier with three separate NBM classifiers for (1), (2), and (3) [VOTE]

Each condition was measured through 10-fold cross-validation which splits the data set into ten different folds (9/10 training vs. 1/10 testing). Each cross-validation run was further seeded with ten different seeds. All reported scores (unless stated otherwise) represent averages from the 10x10-fold cross-validation runs. Three separate baselines are given to reflect a classifier that always outputs a given polarity.

The results from a three-way (POS vs. NTR vs. NEG) classification condition are given in Table 3.

5.2 Evaluation Measures

A large number of different evaluation measures can be used to characterise the performance of the classifiers and the features used, each of which highlights a different evaluative aspect.

Accuracy. The first measure set targets the standard notion of 'accuracy' used in traditional factual classification tasks encompassing **Accuracy**, **Precision**, **Recall**, **F-Score**, and **SAR** measures. For these, individual pairwise polarity conditions (POS vs. NOT-POS, NTR vs. NOT-NTR, NEG vs. NOT-NEG) were used. In addition, raw percentage accuracies are reported. Although sentiment interpretation can not be said to be '(in)accurate' in the strictest sense of the term, these measures characterise the overall behaviour of our classifiers in a useful way.

Agreement. The second set of measures focuses on different levels of agreement and correlation between human sentiment judgements and our classifiers by calculating chance-corrected ternary (POS vs. NTR vs. NEG) rates based on the standard **Kappa** *k*, **Pearson**'s *r* product moment correlation coefficient, **Spearman**'s ρ rank order correlation coefficient, and **Krippendorf**'s α reliability coefficient measures[12].

Error Types. The inter-annotator agreement levels point towards increased ambiguity with NTR polarity due to differing personal degrees of sensitivity towards neutrality/objectivity. Not all classification errors are then equal for classifying a POS case as NTR is more tolerable than classifying it as NEG, for example. We found it useful to characterise three distinct **error classes** or disagreements between human *H* and algorithm *A*. FATAL errors ($H^{(\alpha)}A^{(\neg\alpha)}\ \alpha\in\{+\ -\}$) are those where the non-neutral polarity is completely wrong: such errors affect the performance of a classifier adversely. GREEDY errors ($H^{(N)}A^{(\alpha)}\ \alpha\in\{+\ -\}$) are those where the algorithm wrongly made a decision to jump one way or the other, displaying oversensitivity towards non-neutral polarities. LAZY errors ($H^{(\alpha)}A^{(N)}\ \alpha\in\{+\ -\}$) indicate that the algorithm chose to sit on the fence and displayed oversensitivity towards NTR polarity. We naturally aim at minimising FATAL errors.

5.3 Discussion

In absolute terms, the scores are modest. However, when compared to the low human ceiling (27.75$\sim$59.69%) (§2.2), they do in fact appear promising. In general, all classifiers easily surpassed the low non-neutral (27.6$\sim$30.75%) and neutral[13] (41.69%) baselines. The standalone performance of the social network (SNA) features was not as effective as we expected. In the light of the very small number of features used (only 23), it is in fact surprising that the SNA features worked at all. The average F-score obtained by the SNA features was low (36.3%) mainly due to low recall for non-neutral sentiment. Their pairwise accuracy rates are more favourable as they show that the SNA features are not making random non-neutral predictions. When larger networks are incorporated in the future, social network features can be expected to offer important supporting evidence in the classification task that we are attempting.

Equally promising is the performance of the sentiment analysis (SA) features - especially considering that they only reflect the sentiment scores of a handful of entities and constitute a relatively

[12] All accuracy and agreement measures were obtained using R (http://www.r-project.org/) with the built-in correlation functions together with the ROCR (http://cran.r-project.org/web/packages/ROCR/) and IRR (http://cran.r-project.org/web/packages/irr) packages.

[13] With the exception of the SNA features.

Table 3: Average 3-way 10-fold cross-validation results

	POS	NTR	NEG
BASELINE ACCURACY (RAW)	27.56	**41.69**	30.75

	SNA	SA	BOW	ALL	STACK	VOTE
	(23)	(951)	(4647)	(5620)	(5620)	(5620)
ACCURACY (RAW)	41.05	46.33	**50.34**	49.70	49.29	49.27
ACCURACY (PAIRWISE)	60.70	64.22	**66.89**	66.47	66.20	66.18
ACCURACY (POS)	68.20	65.88	68.06	69.61	**70.59**	69.20
ACCURACY (NTR)	48.79	55.76	**58.66**	57.08	54.83	56.63
ACCURACY (NEG)	65.10	71.03	**73.96**	72.71	73.17	72.71
PRECISION	39.53	45.85	**50.52**	49.40	49.22	49.43
PRECISION (POS)	37.34	36.32	40.72	**42.87**	42.08	41.92
PRECISION (NTR)	42.69	47.54	**50.32**	48.86	47.31	48.54
PRECISION (NEG)	38.57	53.68	**60.52**	56.48	58.26	57.81
RECALL	37.38	44.35	**48.01**	47.49	45.33	46.47
RECALL (POS)	22.64	31.65	**34.46**	30.58	17.52	30.33
RECALL (NTR)	66.67	58.74	65.41	62.79	**73.50**	67.38
RECALL (NEG)	22.81	42.67	44.15	**49.11**	44.96	41.70
F-SCORE	36.30	44.63	**48.41**	47.72	44.33	46.68
F-SCORE (POS)	28.18	33.82	**37.32**	35.67	24.70	35.18
F-SCORE (NTR)	52.05	52.54	56.88	54.95	**57.56**	56.43
F-SCORE (NEG)	28.66	47.53	51.04	**52.53**	50.73	48.45
SAR	50.43	54.28	**56.87**	56.46	55.81	56.04
KAPPA	8.40	19.62	25.15	**25.52**	22.55	23.38
KRIPPENDORFF	44.56	52.84	54.51	**55.64**	51.38	53.25
PEARSON	12.14	23.35	28.67	**30.81**	28.74	27.80
SPEARMAN	12.12	23.46	28.74	**31.18**	29.37	27.99
FATAL ERRORS	13.15	17.55	16.73	14.69	**10.92**	14.51
GREEDY ERRORS	23.56	32.05	29.03	30.83	**21.78**	26.79
LAZY ERRORS	63.29	**50.39**	54.24	54.48	67.30	58.70

small feature set (only 951). The underlying compositional sentiment parser, from which the SA features stem, is very sensitive towards non-neutral sentiment which resulted in slightly higher FATAL and GREEDY error rates for the SA features (cf. the lowest LAZY error rate). The SA features are on the whole more balanced than the SNA ones.

Considering the much larger feature set, the performance of the holistic bag-of-words (BOW) features was (perhaps unsurprisingly) very strong. The figures from 5620 features suggest that, as features, unigram evidence still reflects the sentiment properties of a blog post more closely than non-lexical evidence, even when it comes to measuring sentiment towards or about a single entity. The BOW features were behaviourally closer to the SA features than to the SNA ones which may be due to the fact that the SA entity features latently represent salient unigrams (e.g. *"Obama"*).

Our hypothesis concerning the leveraging power of the SNA and SA features was confirmed partially as small gains over the BOW model were obtained in some conditions by using the entire feature set (ALL) for a single classifier. This can in particular be seen in the higher agreement/correlation rates and lower amounts of FATAL errors. Stacking and voting amongst the three individual NBM classifiers provided further boosts in some conditions.

Interestingly, all classifiers displayed a general tendency towards faring worse with positive polarity than with negative polarity, especially in precision and recall. All classifiers reached their highest recall and F-scores with neutral polarity. Moreover, the amount of FATAL errors stayed below 17% throughout which suggests that all feature types are generally pointing at the right direction. This correlates with the fact that a full 31.94% of the annotations involved conflicting annotations around neutral polarity (NEUTRAL DISAGREEMENT). The classifiers can accordingly be expected cope better with non-neutral sentiment. When all neutral cases were excluded from the evaluation, a small subset of 90...145 non-neutral test cases was examined in order to verify this. Although the subset is too small to draw any definite conclusions, it can nevertheless shed light on how the classifiers did actually do with the core non-neutral cases which are arguably more important for a sentiment classifier than neutral, objective cases.

Table 4 confirms the expected complications caused by neutral polarity in that non-neutral precision scores go as high as 75.68%. Interestingly, the boost over the BOW features given by the non-lexical SNA and SA features is much clearer in the 2-way condition as the combined ALL features and the STACK and VOTE classifiers all outperformed the BOW classifier. These 2-way scores seem to confirm that the SNA and SA features can indeed be used to leverage shallow unigram features in non-neutral cases.

Table 4: Average 2-way 10-fold cross-validation results

	SNA (23)	SA (951)	BOW (4647)	ALL (5620)	STACK (5620)	VOTE (5620)
ACCURACY (PAIRWISE)	63.16	69.84	73.55	76.14	**77.13**	74.23
PRECISION	63.13	69.14	73.06	75.52	**75.68**	73.51
RECALL	63.10	69.73	73.50	**74.66**	72.51	73.78
F-SCORE	63.07	69.21	73.10	**74.95**	73.42	73.57
SAR	55.20	61.58	65.24	**67.32**	67.29	65.77
KAPPA	26.19	38.62	46.35	**49.99**	47.25	47.19
PEARSON	26.23	38.87	46.55	**50.17**	48.08	47.28

We lastly looked at the most informative features based on the Chi-squared and information gain measures across all 5620 features. Of the top 50 features ranked by the two measures, 29 were entity scores from the SA feature set, with the negative entity score for *"Obama"*, *"Chicago"*, and the positive score for *"rhetoric"* topping the ranks alongside high-ranking unigrams such as *"Wright"*, *"pastor"*, and *"Rezko"*, amongst others. This further confirms the utility of the entity-level SA features for our classification task. The SNA features did not rank high as features, however.

5.4 Future Work

The proposed classification framework and the results obtained in this initial study open up many avenues for future work.

Sentiment Analysis. Although their utility is intuitively appealing, it is unclear how far the capabilities of current deep sentiment analysis tools can could bring the analysis considering how subjective the classification task is so. Regarding the deep general-purpose sentiment parser that we employed, further improvements can however be made by tuning its underlying lexicon towards the political domain. The entity-level sentiment scores that were used in this study can further be boosted by resolving pronoun mentions to their antecedents (cf. the amount of pronouns amongst the top-scoring topical entities in Ex. 2).

Wider *n*-grams. A useful research question is whether longer bi- and tri-grams could improve accuracy beyond what we would expect from further blog corpora. Although longer *n*-grams can rudimentarily model some further linguistic features of political discourse, past research in document-level sentiment analysis suggests that simple binary unigram presence features suffice.

Clustering by Topic. Another approach to analysing political blogs comes in the form of blending social network analysis with entity extraction. For every post in our corpus, we have a list of entities that are mentioned in the post, together with their sentiment labels. It is then possible to extract all individual entities mentioned in each post and build a bipartite network representing the data set. In this case, we have a matrix with columns representing entities and rows representing the individual posts (e.g. a value of 1 at the i^{th} row and j^{th} column of the matrix representation of the network denotes that the i^{th} post mentioned the j^{th} entity). That representation would allow us to observe topic overlaps between different posts which would be analysed by constructing a social network between posts, where an edge between posts exists if the posts have some amount (e.g. 10) of entities in common with each other. Such a inter-document network can help elucidate clusters and communities based on topics - not wholly unlike how we use community detection algorithms as part of the social network analysis to locate groups of bloggers that tend to communicate with each other. In this case, however, sharing large numbers of topics or entities in discussions may mean that specific topics - or at the very least, having topics in common - lead to similar sentiment scores. Unfortunately, carrying out such an analysis at this stage yields very poor results for topic clusters: it appears to be the case that most blog posts have a large number of entities in common with each other which causes standard community finding algorithms to group the entire set of posts together as one large community. We conjecture that this is an artifact of how the blog posts were selected when looking for posts discussing Barack Obama. At this stage, the topical clustering method still merits further research.

Advanced Social Network Analysis. At this stage, only a few algorithms have been investigated to combine the various feature sets. While we could explore richer features based on entity extraction, *n*-grams, and deeper social network analysis discussed above, we would still be developing feature vectors for each case and build classifiers that operate on them. Since the algorithms in Weka are not optimised as such for dealing with social network analysis, one extension to our research is to build algorithms that directly integrate predictions into a social network, rather than using features that reflect the current 'membership-within-community' approach. One can in particular use the outputs of the entity-level sentiment scores or unigram probability estimates and apply them as labels in the social network. The social network itself could then be used to reinforce the labels and see which ones appear realistic based on previous observations.

Dataset. A major challenge with the current initial study was data sparsity. Only 700 posts out of 2.8 million were labelled, making it very difficult to extract useful information from the social network features in the data set. We hypothesise that, as more posts are labelled and one gets a finer-grained picture of how bloggers self-organise themselves into topical and political communities, social network features should become more relevant and important. Another issue that ought to be investigated is the concept of data set shift. The fact that many terms and phrases in the political domain develop and change their (non-)affective connotations over time is a key challenge for any text-based political blog and sentiment analysis framework (especially for shallow classification methods). For example, a term such as *"Alaska"* may have had different affective connotations before and after Sarah Palin was announced as the Republican vice presidential candidate.

6 Conclusion

In this paper targeting entity-centric document-level sentiment classification of political blogs, we presented the results of an initial study that sought to investigate the feasibility of combining linguistic indicators of political sentiment with non-linguistic information obtained from social network analysis. Using crowd-sourced sentiment annotations centered around Barack Obama during the 2008 U.S. presidential election sampled from a large corpus of 2.8 million blog posts, a hybrid machine learning and logic-based framework was employed which relies on standard shallow document classification, deep linguistic multi-entity sentiment analysis and scoring, and social network modelling. The initial results demonstrate the complexity of the task, and point towards the positive effects of learning features that exploit entity-level sentiment scores and social network structure.

Acknowledgements

We would like to thank the *Predictive Modeling Group* at IBM for providing us with the blog data, and Nigel Crook at Oxford University Computing Laboratory for supporting code.

References

[1] Lada Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD 2005)*, pages 36–43, NewYork, NY, USA, August 21 2005.

[2] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

[3] G. Csárdi and T. Nepusz. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695, 2006.

[4] Kathleen T. Durant and Michael D. Smith. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In *Advances in Web Mining and Web Usage Analysis: Proceedings of the 8th International Workshop on Knowledge Discovery on the Web (WEBKDD 2006)*, pages 187–206. Philadelphia, USA, August 20 2007.

[5] Miles Efron. Cultural orientation: Classifying subjective documents by cocitation analysis. In *Style and Meaning in Language, Art, Music, and Design: Papers from the 2004 AAAI Fall Symposium*, Technical Report FS-04-07, pages 41–48. Arlington, Virginia, USA, October 21-24 2004.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[7] E. Hargittai, J. Gallo, and M. Kane. Cross-ideological discussions among conservative and liberal bloggers. *Public Choice*, 134(1):67–86, 2008.

[8] P.Y. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, 2009.

[9] Soo-Min Kim and Eduard Hovy. Crystal: Analyzing predictive opinions on the web. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1056–1064, Prague, Czech Republic, June 2007.

[10] A. Kittur, E.H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the 26th annual SIGCHI conference on Human factors in computing systems*, pages 453–456, Florence, Italy, April 5-10 2008. ACM New York, NY, USA.

[11] Namhee Kwon, Stuart W. Shulman, and Eduard Hovy. Multidimensional text analysis for erulemaking. In *Proceedings of the 7th Annual International Conference on Digital Government Research (DG.O 2006)*, pages 157–166, San Diego, CA, USA, May 21-24 2006.

[12] Kevin Lerman, Ari Gilder, Mark Dredze, and Fernando Pereira. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 473–480, Manchester, UK, August 2008.

[13] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. *Society of Applied and Industrial Mathematics: Data Mining*, 2007.

[14] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*, pages 109–116, New York City, USA, June 8-9 2006.

[15] Robert Malouf and Tony Mullen. Taking sides: User classification for informal online political discourse. *Internet Research*, 18(2):177–190, 2008.

[16] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *An introduction to information retrieval*. Cambridge University Press, 2008.

[17] M. McPherson, L. Smith-Lovin, and J.M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[18] Karo Moilanen and Stephen Pulman. Multi-entity sentiment scoring. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2009)*, pages 258–263, Borovets, Bulgaria, September 14-16 2009.

[19] Tony Mullen and Robert Malouf. Preliminary investigation into sentiment analysis of informal political discourse. In *Computational Approaches to Analyzing Weblogs: Papers from 2006 AAAI Spring Symposium*, pages 159–162. Stanford, California, USA, March 27-29 2006.

[20] Bo Pang and Lillian Lee. *Opinion Mining and Sentiment Analysis*, volume 2 of *Foundations and Trends in Information Retrieval*. Now Publishers Inc, 2008.

[21] MF Porter. An algorithm for suffix stripping. *Program*, 14(1):3, 1980.

[22] R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 254–263, Honolulu, Hawaii, 2008.

[23] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 327–335, Sydney, Australia, July 22-23 2006.

[24] M. Tremayne, N. Zheng, J.K. Lee, and J. Jeong. Issue publics on the web: Applying network theory to the war blogosphere. *Journal of Computer-Mediated Communication*, 12(1):290–310, 2006.

[25] Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Exploring the characteristics of opinion expressions for political opinion classification. In *Proceedings of the 9th Annual International Conference on Digital Government Research, Partnerships for Public Innovation (DG.O 2008)*, volume 289, pages 82–91, Montreal, Canada, May 18-21 2008.