Affinity Filtering
A Novel Approach to Social Recommendation

Riley Kidd

A subthesis submitted in partial fulfillment of the degree of
Bachelor of Software Engineering at
The Department of Computer Science
Australian National University

October 2012

Typeset in Palatino by TEX and LATEX 2ε.

Except where otherwise indicated, this thesis is my own original work.

Riley Kidd
22 October 2012

Abstract

Social networks such as Facebook allow users to create a rich and verbose profile composed of both user specific interactions (such as comment and message passing, tags, likes) and user preferences (such as favourite movies and music, group memberships, page likes). These distinct components of a users profile can be leveraged to gauge their affinity with certain links and ultimately predict their explicit like preferences.

The goal of this thesis is to decipher which of these aforementioned affinity measures are truly predictive of a users like preferences.

The success of our predictions are evaluated using the machine learning algorithms of *Naive Bayes*, *Logistic Regression* and *Support Vector Machines*, results are compared to previous work using the state of the art social collaborative filtering technique of *Social Matchbox* as a baseline. The data set is sourced from a set of over 100 Facebook users and their interactions with over 39,000+ friends during a four month period.

Our analysis has shown that user interactions in themselves are not highly predictive of user likes, while user preferences are. We conclude by analysing a combination of the most predictive user preference measures, offer a summary of our work to date and propose recommendations for additional research in this area.

Contents

Chapter 1

Introduction

The Internet is becoming a network of people, providing a myriad of expanding social information and user driven content. Social presence on the web is continually expanding. With the emergence of services such as Facebook, Myspace, LinkedIn, Twitter and Google+, what defines a user and their online *user interactions* (such as comment and message passing, tags, likes) and *user preferences* (such as favourite movies and music, group memberships, page likes) is an expanding graph of rich social content.

From this premise, the ultimate question we wish to address in this thesis is: How can we leverage this user information to decipher which *user interaction* or *user preference* affinity features are most predictive of user likes?

We address this question by comparing and contrasting these different potential affinity relationships in our data against appropriate baselines and ultimately offer a combination of features which offers our best solution to the question posed above.

In this chapter we will outline the objectives of this research, summarise the contributions made and provide an outline for the remaining chapters.

1.1 Objectives

One issue present in this Facebook paradigm is discovering whether a user doesn't like an item, a users Facebook feed is comprised of activity between their friends, content, groups, etc giving an enormous scope of potential feed items. Facebook will only show feed items for users who have recently interacted with using their *Edge-Rank* [Sanghvi and Steinberg 2010] algorithm.

While many Facebook users have a friend count which is close to the human real word limit, known as the Dunbar number [Hill and Dunbar 2003], the *Edge-Rank* algorithm ensures user interactions are focused on a much smaller subset of their friends. Additionally, given the rate of posting, these top feed items are only displayed for a short amount of time. Coupled with the fact that Facebook allows users to explicitly like an item, but not dislike it - distinguishing between what a user does and does not like becomes difficult.

The primary objective of this thesis is to compare and contrast differing potential affinity features across *user interactions* and *user preferences*. Using state of the art ma-

chine learning concepts of *Naive Bayes* (NB), *Logistic Regression* (LR) and *Support Vector Machines* (SVM) compared with our appropriate baselines of *Social Matchbox* (SMB) and *Constant Classifiers*. With the primary aim of discovering which affinity features are most predictive or user likes.

Based on the insight that social inuence can play a crucial role in a range of behavioural phenomena [Granovetter 1978; Watts and Strogatz 1998] and that positive social annotations on search items add perceived utility to the worth of a result [Pantel and Haas 2012] we will also test using an exposure hold out technique, where data is only tested if some friend has already liked that item. Hence the need to undertake an analysis of the effect of exposure on our affinity features.

Finally, we will assess and compare the effect of combining successful individual affinity features found during our analysis.

1.2 Contributions

Our specific contributions made during this thesis show:

- Both *interactions* and *incoming \outgoing messages* posted between users are not more predictive than previously used SMB techniques.
- Each *user preference* affinity of *favourites* (such as favourite movies, music), *group* memberships (such as Australian National University and Students in Canberra) and *page* likes (such as Google Chrome and The Simpsons) outperformed our baselines.
- Combining both affinity types of *user interactions* and *user preferences* with an exposure limit resulted in a substantial improvement over previous techniques as the exposure increases.
- Combination of the advantageous affinity features briefly mentioned above gives the best results in our analysis.

Overall, we provide a methodology that improves upon previous work and offers an approach to combine predictive affinity features.

1.3 Outline

The remaining chapters in this thesis are organised as follows:

- **Chapter 2**: We first outline appropriate background information for the reader. Including information pertaining to the source of our data set, mathematical notation used throughout this thesis, previous work in this area and our research approach and methodology.

- **Chapter 3**: In this chapter we discuss different affinity features for *user interactions* and the results of applying these features to NB, SVM and LR in comparison with our baselines.
- **Chapter 4**: A similar affinity feature analysis as above is applied, however the features we utilise are for *user preferences*.
- **Chapter 5**: In this chapter we discuss the effect of combining different affinity features based on results gained in the previous sections and propose an ideal affinity feature hybrid.
- **Chapter 6**: Finally, we draw the work done throughout this thesis to a conclusion and offer avenues for future work in this area.

With all chapters combined, this thesis represents a novel approach to exploiting and analysing *user interactions* and *user preferences* affinity relationships to ascertain which features are most predictive of user likes and present an approach of combining these useful feature components into an effective classification paradigm.

Chapter 2

Background

In this chapter, we define the social network Facebook central to this study, the source of our data set, notation used throughout this thesis, our choice of classification algorithms and our testing approach and methodology.

2.1 Facebook

Facebook is the largest and most active social media service in the world (as of September 2012 it had more than 1 billion active users [**?**]). Facebook users can create a profile containing personal preferences and information including their favourite music, favourite movies, inspirational people, interests, age, birthday, etc and have friendships and interactions between other users.

The four main interactions between users are posts (posting something on a friends' wall), tags (being mentioned in a friends post or comment), comments (written data on a post) and likes (clicking a like button if a user likes a post or comment). The mediums for these interactions are across links (some URL), posts (some Facebook post), photos (some uploaded Facebook photo) and videos (some uploaded Facebook video).

Given the enormous scope of interaction and preference information available about each user, NICTA have developed an app capable of tracking and recording all pertinent user information. This app will be discussed in the following section.

2.2 Data Set

NICTA developed a Facebook app named *LinkR*. [1] This app collected information about users, their interactions and preferences as well as a subset of available information about their friends. The app tracked and stored this information for over 100 app users and their 39,000+ friends over a 4-month time period.

Exhaustive interaction and profile information could not be recorded for the app

[1] The main developer of the LinkR Facebook App is Khoi-Nguyen Tran, a PhD student at the Australian National University.

users friends and as such all analysis performed in this thesis was carried out exclusively on app users for whom we have full interaction and profile data.

The table below summarises the interactions data collected from both app users and their friends which is used during our subsequent analysis.

App Users	Posts	Tags	Comments	Likes
Wall	36,539	7,711	18,266	15,999
Link	5,304	-	5,757	6,566
Photo	4,933	28,341	8,677	8,612
Video	245	2,525	1,687	843
App Users and Friends	**Posts**	**Tags**	**Comments**	**Likes**
Wall	4,301,306	1,215,382	3,122,019	1,887,497
Link	678,612	-	693,930	995,214
Photo	1,268,816	9,620,708	3,431,321	2,469,859
Video	59,244	904,604	486,677	332,619

Table 2.1: Data records for interactions between users. Rows are the type of interaction, columns are the medium of interaction.

2.3 Notation

The mathematical notation utilised during this thesis is outlined below.

- N users U with an I-element user feature vector X where $X \in \mathbb{R}^I$ (alternatively if a second user is needed $Z \in \mathbb{R}^I$) and the length and components of I are uniquely defined for each affinity feature.
- A set of items V.
- A friend function $Friend_{u,z}$ which is *True* when users u and z are friends.
- A liked function $Liked_{u,v}$ which is *True* when user u likes item v.
- A relationship R between users where $R_{i,u}$ is uniquely defined for each affinity feature.
- An alters set A for each user u item v pair, based on some relationship R between other users z. Where $a_{u,v,r} = \{z | R_{u,z} \wedge Likes_{z,v}\}$.

This alters set can be visualised in the figure below:

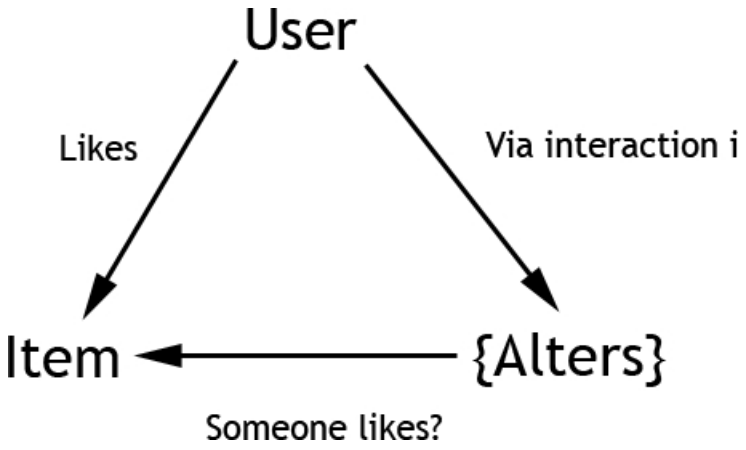

Figure 2.1: Alters paradigm. A user u likes some item m, a relationship $R_{u,z}$ is defined via some affinity i uniquely defined for each affinity feature, to create our set of alters A.

- An exposure E where $E_{u,v,z} = \sum_{z}^{N} Friend_{u,z} \wedge Liked_{z,v}$ where this exposure can be limited by some k with the condition $E_{u,v} >= k$.

  This exposure can be visualised in the figure below:

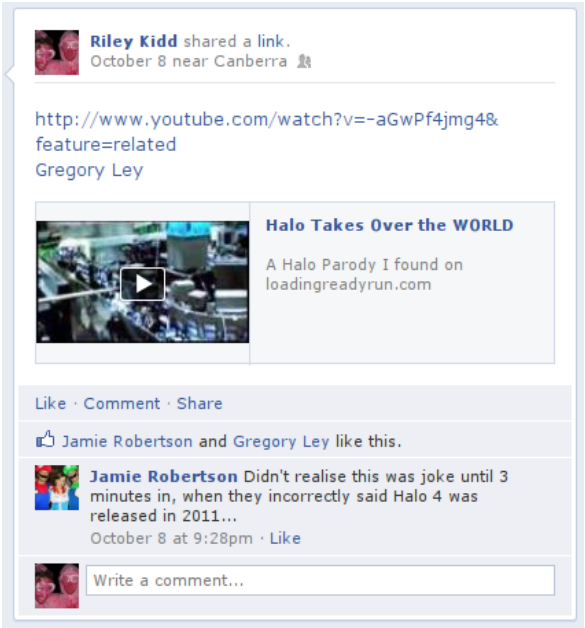

Figure 2.2: Here we see an example of a link posted to a friends wall, which has subsequently been liked by two friends z. This demonstrates an exposure of 2 for this link m.

- A data-set D comprised of $D = \{(u, v, x) \rightarrow y\}$ where $u \in U$, $v \in V$ and the binary response $y \in \{0, 1\}$ where 0 represents a dislike and 1 represents a like.

2.4 Affinity Features

Given the vast amount of potential affinity features available on Facebook, we need to break our features down into separate, distinct groups for testing purposes. Our two main groups will be *user interactions* and *user preferences*.

The individual components of these groups are displayed below: *User interactions*:

- *Interactions* : Posts, Tags, Likes, Comments.
- *Outgoing Messages* : Messages sent to other users.
- *Incoming Messages* : Messages received from other users.

User preferences:

- *Demographics* : Age, Gender, Location.
- *Favourites* : Activities, Books, Athletes, Teams, Movies, Music, Sports, Television, Movies, People, Interests.
- *Groups* : All groups a user has joined.
- *Pages* : All pages a user has liked.

Each of these affinity features will be discussed in detail under their separate sections of this thesis. During our analysis we will compare the predictiveness of each of these affinity features individually, as well as in combination.

2.5 Previous Work

Two general approaches to prediction in a social context are *content-based filtering* (CBF) [Lang 1995] which exploits item features based on items a user has previously liked and *collaborative filtering* (CF) [Resnick and Varian 1997] which exploits the current users preferences as well as those of other users.

Previous work defined the term *social* CF (SCF) [Noel 2011] which augments traditional CF methods with additional social network information, the results of this previous work and analysis using live user trials came to the conclusion that the approach of SMB provided the best results for this data set and as such will be used as a baseline in this thesis.

These methods of CBF, CF and SCF result in a user gaining some similarity measure between other users, while the affinity features we explore during this thesis are based on explicit *user interaction* and *user preference* features and result in different models and predictions based on our choice of feature selection.

2.6 Training and Testing

All evaluation is applied using 10 fold cross validation wherein the data is partitioned into 10 complementary subsets, 80% of these subsets are used for training while the remaining 20% are used for testing.

The training and testing process is repeated 10 times for each set of fold data. These results are then averaged to produce our estimates and standard error. The benefit of this method over repeated sub-sampling is all data points are used for both training and validation.

2.7 Classification Algorithms

Each classification algorithm used in this thesis is passed the training data for each fold as outlined above. The classifier builds a model representation of the data and applies this model to the test set to classify each test item into either a like or a dislike.

All affinity feature analysis carried out in this thesis will be performed on the following classification algorithms:

2.7.1 Constant

The constant predictor returns a constant result irrespective of the feature vectors selected. Namely, this predictor returns 0 (false) regardless of the affinity feature represented by X. The most common result in our data set is *False* and hence the *False* predictor is displayed in all analysis, tables and graphs in this thesis.

2.7.2 Social Match Box

SMB is an extension of existing SCF techniques [Yang et al. 2011; Cui et al. 2011] which constrain the latent space to enforce users who have similar preferences to maintain similar latent representations when they interact heavily.

SMB uses the social regularization method which incorporates user features to learn similarities between users in the latent space which allows us to incorporate the social information of the Facebook data [Noel 2011].

This objective component constrains users with a high similarity rating to have the same values in the latent feature space, which models the assumption that users who are similar socially should also have similar preferences for items.

2.7.3 Naive Bayes

Naive Bayes is a basic probabilistic classifier which involves applying Bayes' theorem using strong conditional independence assumptions between each feature in X. During training each element i in the feature vector X is devised to contribute some evidence that this x_i belongs to either a like or dislike classification, during testing the

class with the highest probability when applied to the model is the classification predicted.

With feature vector $f \in \mathbb{R}^F$ derived from $(x, y) \in D$, denoted as $f_{x,y}$. NB learns a conditional model of the form $p(C|F_1, \ldots, F_n)$ over a dependent class variable C conditioned on the feature variables $F_1, \ldots, F_n$. Applying both Bayes' rule and conditional independence assumptions the model can be rewritten as $p(C|F_1, \ldots, F_n) = p(C) \prod_{i=1}^{n} p(F_i|C)$.

Classification of our test vector is achieved by choosing the most probable class of either like (1) or dislike (0).

$$\text{classify}(f_1, \ldots, f_n) = \underset{c \in \{1,0\}}{\text{argmax}}\ p(C = c) \prod_{i=1}^{n} p(F_i = f_i|C = c).$$

2.7.4 Logistic Regression

Logistic Regression directly estimates parameters based on the training data assuming a parametric form of the distribution. LR predicts the odds of a feature vector X being either a like or a dislike by converting a dependent variable and one or more continuous independent variable(s) into probability odds.

The probability p_i is modelled using a linear predictor function $l(i)$, the linear predictor function of a particular point d is written as $l(i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_M x_{m,i}$, where each data point d is associated with an explanatory feature vector X and $\beta_0, \ldots, \beta_M$ are regression co-efficients indicating the relative effect of a particular explanatory variable $x_{m,i}$ on the prediction.

The probability of a particular outcome is linked to the linear prediction function, $\text{logit}(\mathbb{E}[Y_i|x_{1,i}, \ldots, x_{m,i}]) = \text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_M x_{m,i}$ Where the class of either dislike (0) or like (1) with the higher probability is the prediction made.

The LR implementation used during this thesis is *LingPipe* [Alias-i. 2008. LingPipe 4.1.0. http://alias-i.com/lingpipe (accessed October 1 2011].

2.7.5 Support Vector Machine

The *Support Vector Machine* is a supervised learning machine based on a set of basis functions which help construct a separating hyperplane between data points. Training involves building the relevant hyperplanes which can then be used for testing. Each data point is classified as a like or dislike depending on which side of the hyperplane it falls.

With feature vector $f \in \mathbb{R}^F$ derived from $(x, y) \in D$, denoted as $f_{x,y}$. A linear SVM learns a weight vector $w \in \mathbb{R}^F$ such that $w^T f_{x,y} > 0$ indicates a like classification of $f_{x,y}$ and $w^T f_{x,y} \leq 0$ indicates a dislike classification.

The SVM implementation used during this thesis is *SVMLibLinear* [Chang and Lin 2011].

2.8 Evaluation Metrics

When evaluating the success of each affinity feature at correctly classifying an item, the following metrics have been analysed.

- A *true positive* (TP) prediction refers to when the prediction correctly identifies the class as true.
- A *false positive* (FP) occurs when the prediction is true, but the true class was false.
- A *false negative* (FN) occurs when the prediction is false but the actual class is true.

These definitions can be visualised using the table below. Where:
y represents the true class value $y \in \{0, 1\}$: *actual*
$\hat{y}$ represents the class prediction $\hat{y} \in \{0, 1\}$: *prediction*.

		y	
		T	**F**
$\hat{y}$	**T**	TP	FP
	F	FN	TN

Table 2.2: Actual and prediction comparison table.

Accuracy relates to the closeness to the true value. In the context of our results, the accuracy refers to the number of correct classifications divided by the size of the data set.

$$\text{accuracy} = \frac{\text{number of correct classifications}}{\text{size of the test data set}}$$

Precision relates to the number of retrieved predictions which are relevant. In the context of our results, the precision refers to the number of TP predictions divided by the sum of the TP and FP predictions.

$$\text{precision} = \frac{\text{number of TP}}{\text{number of TP + number of FP}}$$

Recall refers to the number of relevant predictions that are retrieved. In the context of our results, recall refers to the number of TP predictions divided by the sum of the TP and FN predictions.

$$\text{recall} = \frac{\text{number of TP}}{\text{number of TP + number of FN}}$$

The f-score combines and balances both precision and recall and is interpreted as the weighted average of both precision and recall.

$$\text{f-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

The main metric we use for analysis, tabulation and graphing in this thesis is accuracy.

Chapter 3

User Interactions

This chapter is dedicated to analysing the different *user interaction* features present in Facebook.

The *user interactions* we examine in this thesis can be broken down into two distinct groups:

- **Interactions** : Posts, Tags, Likes, Comments between users.
- **Messages** : Both outgoing and incoming messages sent between users.

These interactions give implicit networks of friendships, previous methods [www] have claimed if people interact frequently they will like the same things, in this section we break this idea down into the smaller implicit overlaps of these interactions.

3.1 Interactions

Interactions between users in Facebook can be summarised under the following categories:

- **Direction**: The manner an interaction is received, either *incoming* for example where a message is posted to some user or *outgoing* where some user posts a message to another user. Interaction directionality has been shown to be highly reflective of user preferences [Saez-Trumper et al. 2011].
- **Modality**: The medium some user employs to interact with another user via either *links*, *posts*, *photos* or *videos*.
- **Type**: The style some user employs to interact with another user via either *comments*, *tags* or *likes*.

For *user interactions* each user u and feature vector x is defined as the cross product of the above components where:

$$I = \{Incoming, Outgoing\} \times \{Posts, Photos, Videos, Links\} \times \{Comments, Tags, Likes\}$$

The alters set for each i is conditioned by the relationship R, where:

$$R_{i,u} = \{z | Interacted_{i,u,z}\}$$

In this case the *Interacted* function returns all users z who have interacted with user u via the current interaction i.

Applying this affinity feature to our classification algorithms we obtain:

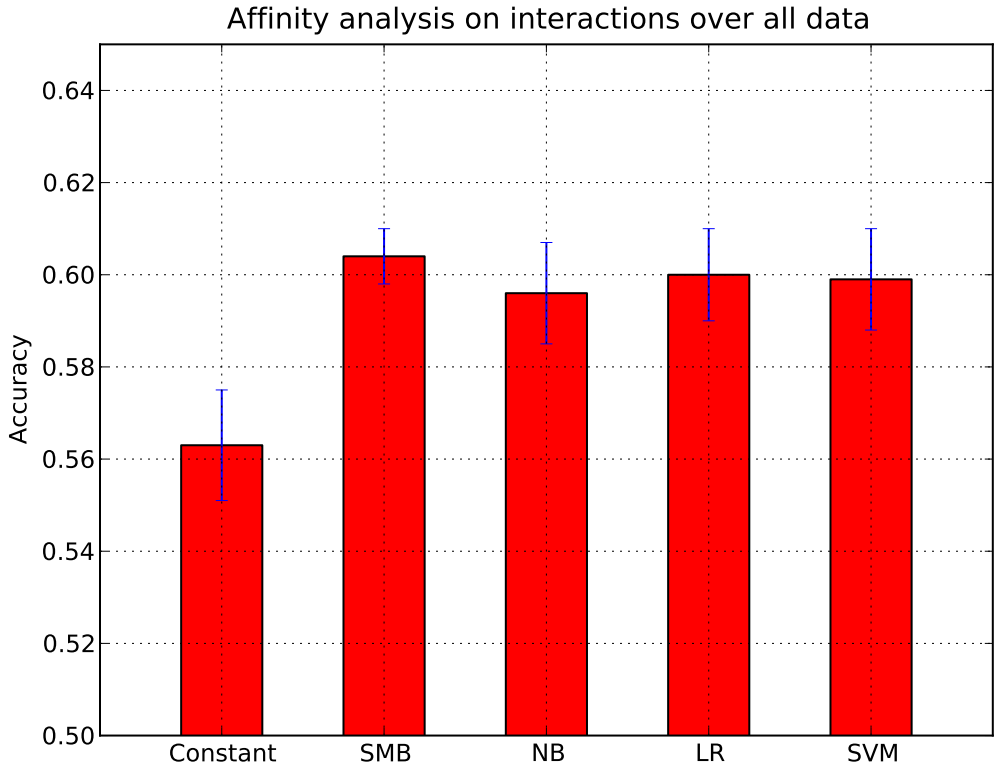

Figure 3.1: Accuracy results using *user interactions* against all data. *User interactions* do not outperform our baselines.

User interactions in themselves are not more predictive then our SMB baseline. One reason for this result could be we can not track information passing outside of Facebook, users who frequently interact could be real world friends and hence share information via email or word of mouth and not over Facebook.

Comparing *user interactions* against an exposure across k we obtain:

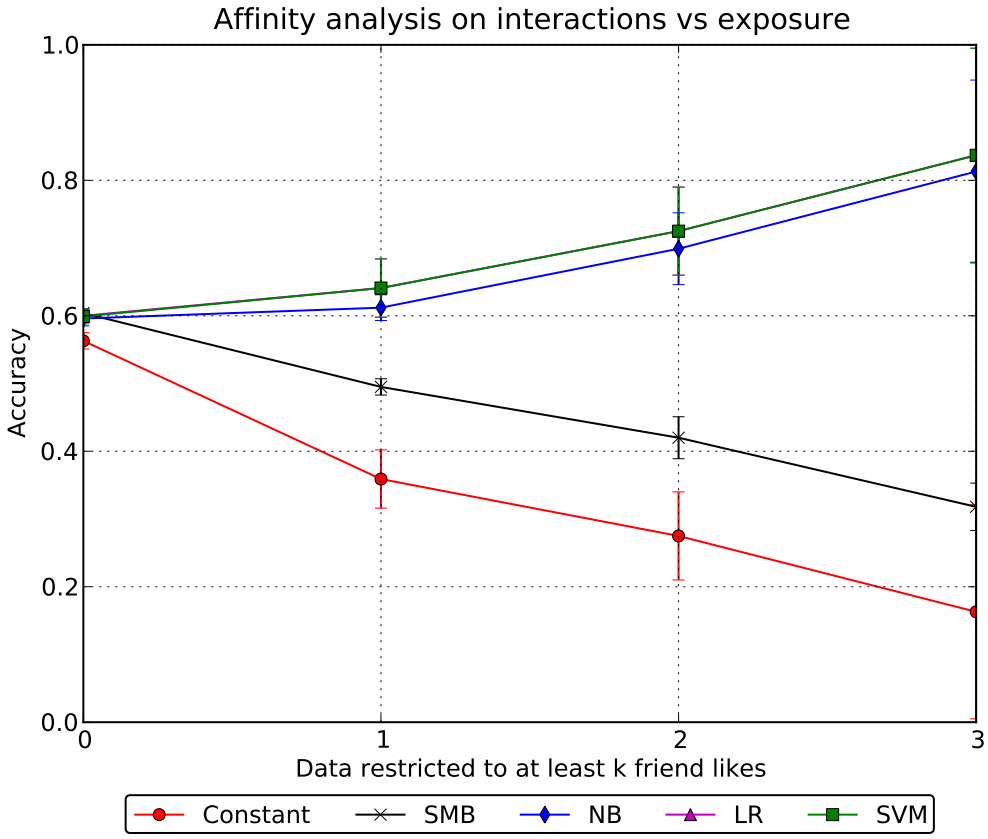

Figure 3.2: Accuracy results against exposure using *user interaction* features. *User interactions* provide a drastic improvement over our baselines as k increases, suggesting SMB is not always the best classifier. This demonstrates the intuitive assumption that *user interactions* can not improve prediction when these interactions do not exist between users. Note in this case LR and SVM both learnt the same result.

Our comparison has shown that as our data is restricted across an exposure, the performance of our classifiers improves. This implies that for *user interactions* simply having one user liking an item is enough to improve upon our baselines. This is intuitively correct as our classifiers can not learn when interacts do not exist between users.

3.2 Conversation

The next *user interactions* we compare are messages passed between users.

These messages can be broken down based on their directionality, either *outgoing* which are messages sent to other users or *incoming* which are messages received from other users.

Based on our data set, the most common words occur with a high frequency and are outlined in the table below:

Rank	Word	Frequency
1	:)	292,733
2	like	198,289
3	good	164,387
4	thanks	159,238
5	one	156,696
6	love	139,939
7	:p	121,904
8	time	106,995
9	think	106,459
10	see	103,690
11	nice	99,672
12	now	94,947
13	well	92,735
14	happy	84,381
15	:d	83,698
16	much	78,719
17	oh	77,321
18	yeah	76,564
19	back	76,032
20	great	70,514
21	going	70,447
22	still	68,245
23	new	67,430
24	day	65,579
25	come	63,837
26	;)	62,936
27	year	61,771
28	look	60,608
29	yes	59,774
30	want	59,514
31	tag	58,633
32	hahaha	57,448
33	also	56,414
34	need	55,921
35	make	54,949
36	sure	54,395
37	thank	54,112
38	people	53,211
39	miss	53,182
40	guys	52,855

Table 3.1: Top conversation content data for all users. We see very common words and online expressions have a high frequency in our data set. Additionally highly emotive and sentimental words are very common, implying these interactions occur between *real* friends.

For messages each user u and feature vector x is defined as the cross product of:

$$I = \{Incoming, Outgoing\} \times \{MessagesSize\}$$

Where the optimal *messages size* J is defined for each directionality and classifier. The alters set for each i is conditioned by the relationship R, where:

$$R_{i,u} = \{z | Messaged_{i,u,z,j}\}$$

In this case the *Messaged* function returns all users z who have messaged user u via the current messaging direction i with the word at index j.

3.2.1 Outgoing

The first issue is to determine the most predictive number of *outgoing* words j for use by our classifiers. Given the expansive size of potential messages and memory constraints in the testing environment we decided to test within a range of $\{100 - 1000\}$ with an incremental step size of 100 for each test.

The results of testing based on differing sizes of *outgoing* words can be seen below:

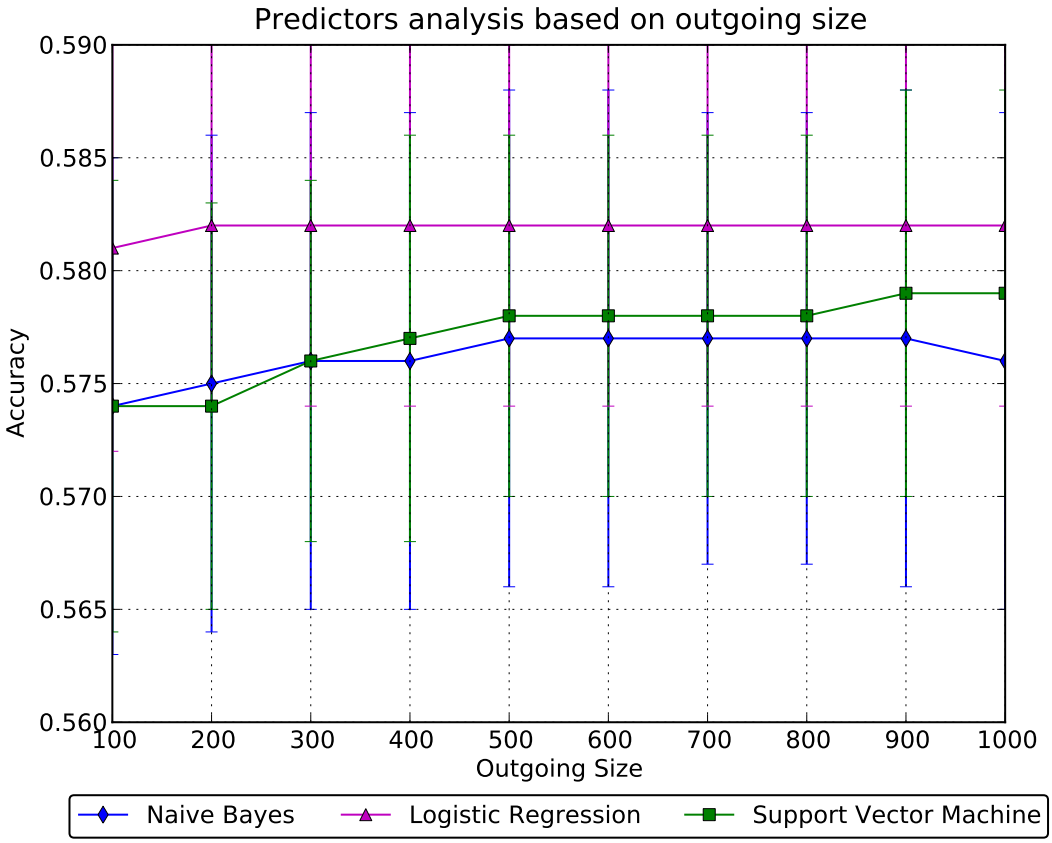

Figure 3.3: Accuracy results for different *outgoing* words sizes. Best performance can be found using LR with a relatively small word size of only 200.

The most predictive *outgoing* words sizes j for each of our classifiers are:

- **Naive Bayes**: 500
- **Logistic Regression**: 200
- **Support Vector Machine**: 900

Using the most predictive word sizes j for each of our classifiers and building our feature vector as defined above we obtain:

These results do not show an improvement over our baselines and are only a marginal improvement over the *constant* baseline.

A possible reason for this could be due to the commonality of the words being tested. Highly common and frequently used words would result in poor predictive tendencies.

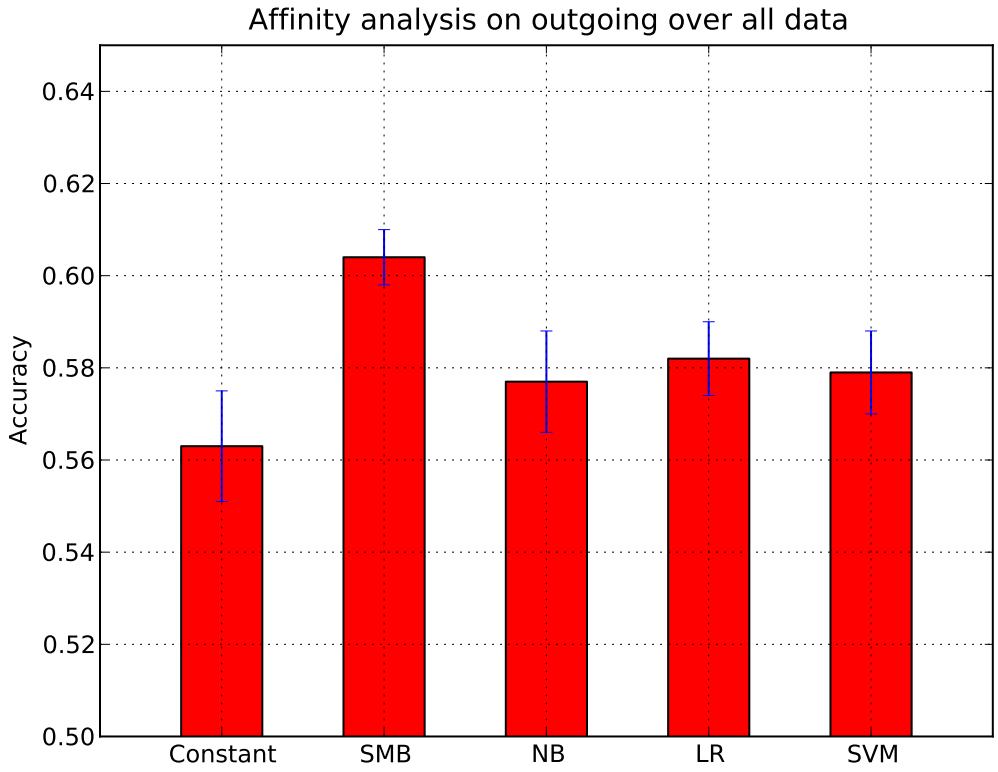

Figure 3.4: Accuracy results using the *outgoing* words features. *Outgoing* words are clearly less predictive then *user interactions*.

Comparing *outgoing* words against exposure we obtain:

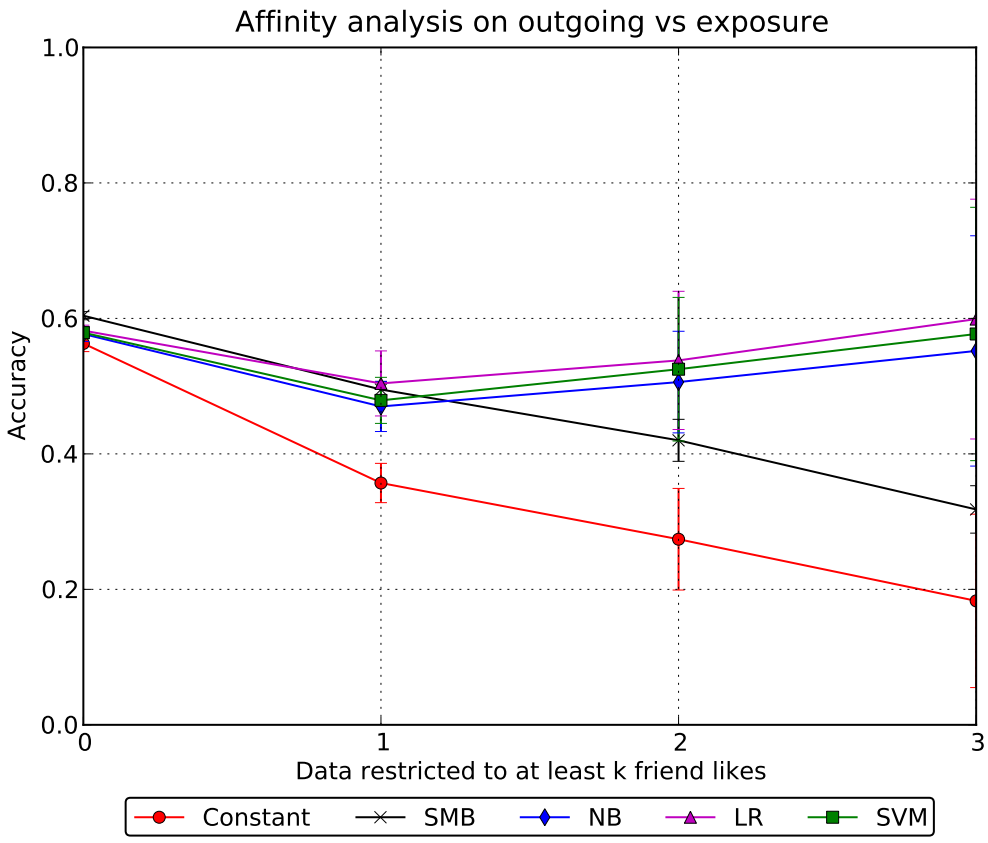

Figure 3.5: Accuracy results against exposure using the *outgoing* words feature. *Outgoing* words predictiveness improve as k increases, but are still less predictive even in the case for $k = 3$ when compared with SMB when $k = 0$.

The exposure offers an increase in the predictiveness of the *outgoing* words feature. However even when $k = 3$, this feature does not outperform SMB when $k = 0$. Indicating that *outgoing* words are not a predictive feature in our data set.

3.2.2 Incoming

Similarly for *incoming* words we need to discover which is the most predictive word size j for use by our classifiers, using the same methodology as described above for *outgoing* words we obtain the following graph:

The most predictive *incoming* words sizes j for each of our classifiers are:

- **Naive Bayes**: 300
- **Logistic Regression**: 100
- **Support Vector Machine**: 1000

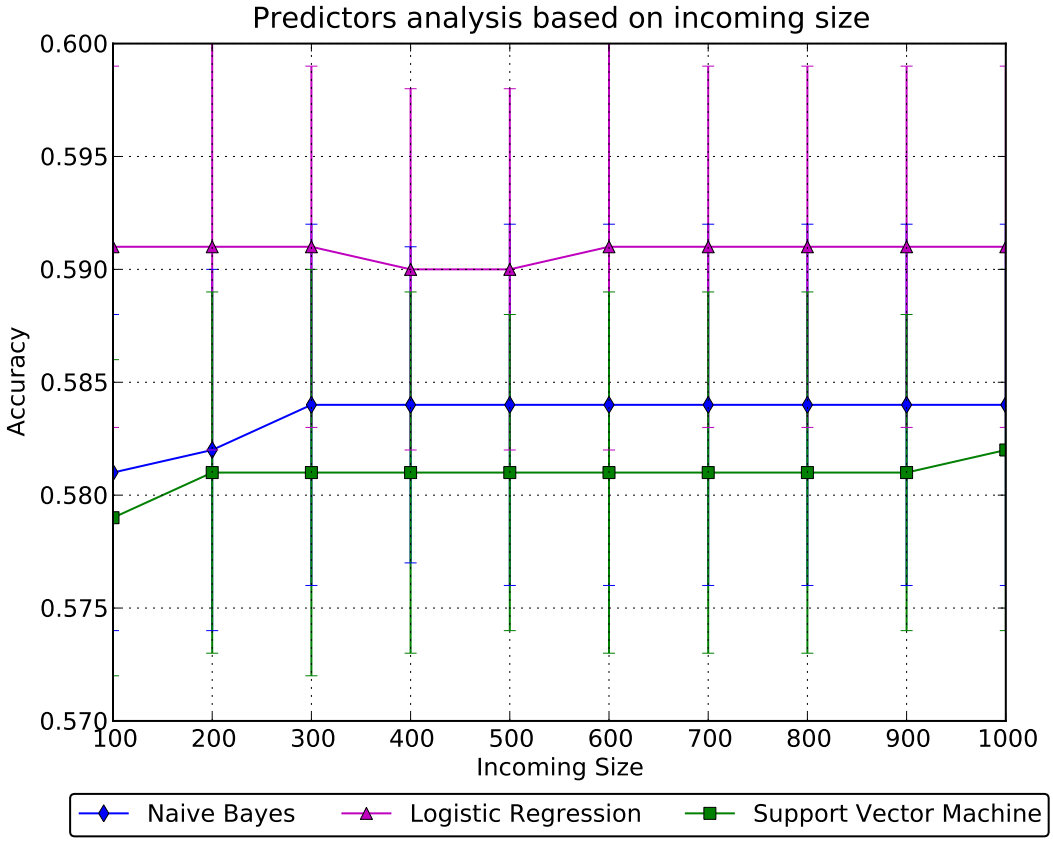

Figure 3.6: Accuracy results for different *incoming* words sizes. *Incoming* words are more predictive then *outgoing* words, but follow the similar trend that small sizes offer close to optimal predictions for this feature.

Using these most predictive word sizes for each of our classifiers we obtain:

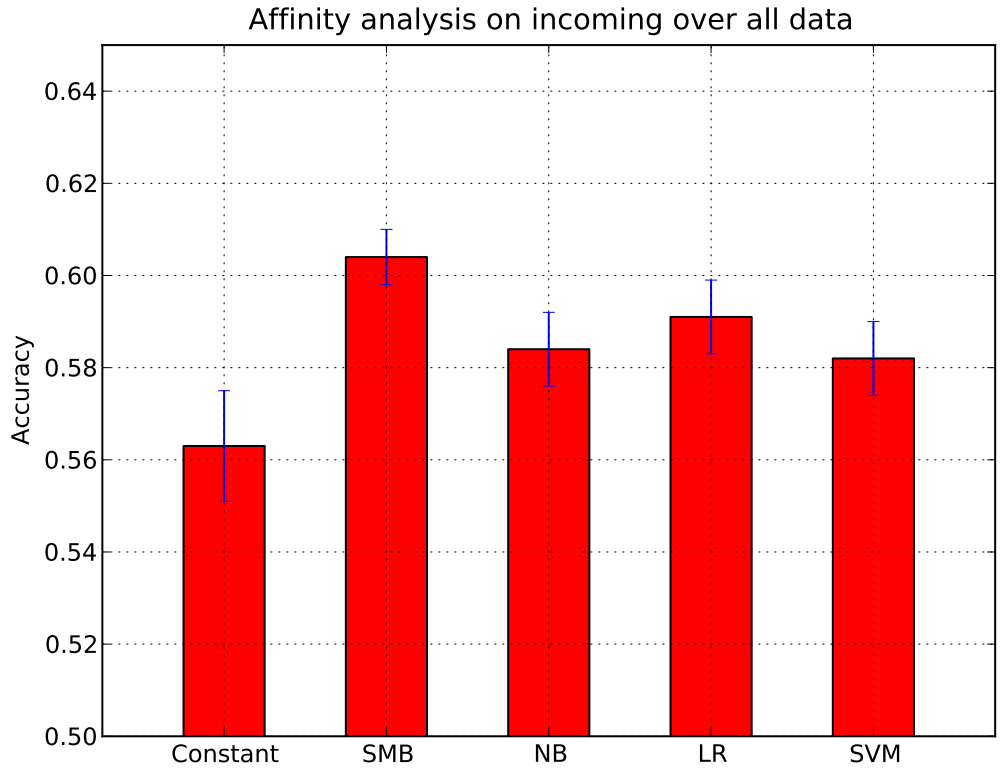

Figure 3.7: Accuracy results using the *incoming* words features. *Incoming* words are a stronger predictor then *outgoing* words, however they are still a weaker predictor then *user interactions*.

Similarly to *outgoing* words, *incoming* words themselves are not more predictive in comparison to our baselines.

Comparing *incoming* words against exposure we obtain:

Similarly, *incoming* words improve upon our baselines as k increases, however this predictive increase is negligible in comparison with $k = 0$ and hence *incoming* words are not predictive of user likes.

Incoming words show an improvement over *outgoing* words, however neither outperforms *user interactions*.

3.3 Conclusion

Throughout this chapter we have explored the different interaction types available between users on Facebook.

We have found that words, irrespective of their directionality do not assist in im-

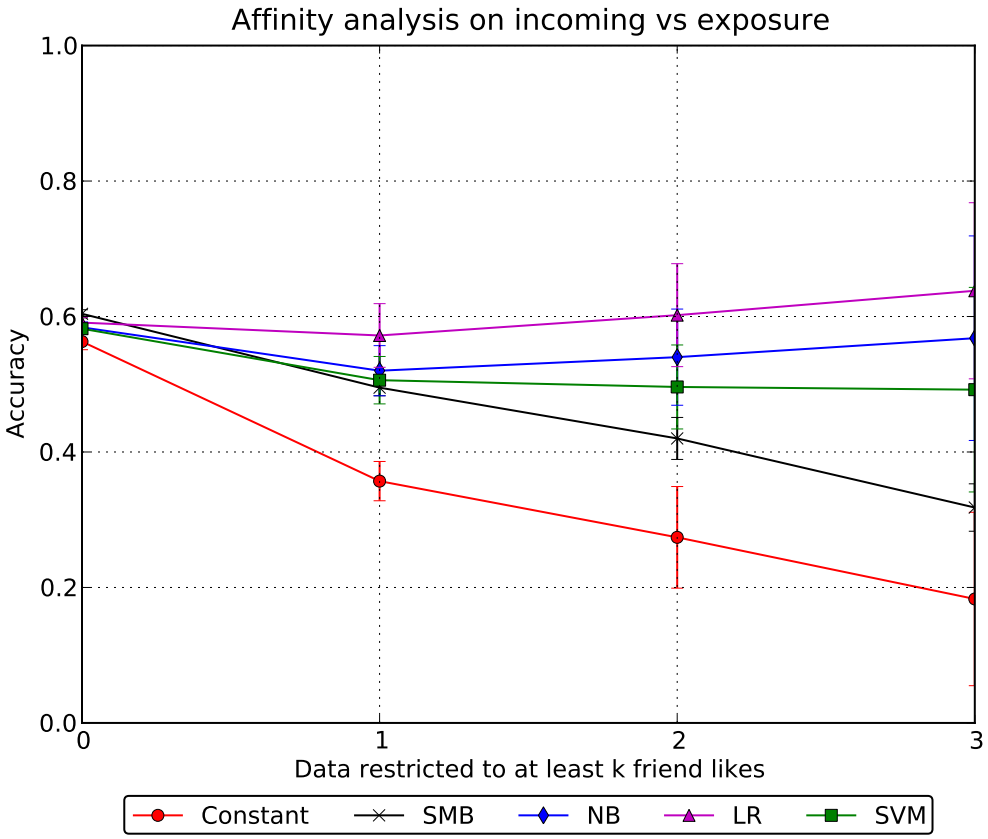

Figure 3.8: Accuracy results against exposure using *incoming* words features. *Incoming* words predictive accuracy improves as k increases, but are still less predictive then *user interactions*.

proving predictions. [Anderson et al. 2012] concluded that it is less important what users say, then who they interact with, which we also found in our results. Additionally, it has been shown for *user interactions* outgoing interactions are more important [www], while our results have found that for *words*, *incoming* are more important.

Our results have shown, that for *user interactions* to improve upon baseline prediction it is enough for some user to have previously liked the item, this allows our classification methodology to improve predictiveness as k increases because interactions exist between these users and our classifiers can learn from them.

Chapter 4

User Preferences

In this chapter we will compare the predictiveness of applying different types of *user preferences* as our feature vector.

The *user preferences* we examine during this thesis are:

- **Demographics** : Age, Gender, Location.
- **Favourites** : Activities, Books, Athletes, Teams, Movies, Music, Sports, Television, People, Interests.
- **Groups** : All groups a user has joined.
- **Pages** : All pages a user has liked.

4.1 Demographics

The *demographics* data we are interested in includes:

- **Age**
- **Birthday**
- **Location**

Below we will give a basic analysis of the *demographics* data in our data set.
Gender breakdown:

Male	Female	Undisclosed
85	33	1

Table 4.1: Gender breakdown for app users. We see a strong male bias.

Despite this clear male bias [Ugander and Marlow 2011] found that in a social setting, there are no strong gender homophily tendencies. Hence the male skew should not negatively affect our results. Additionally [Backstrom et al. 2011] have shown that different genders have differing tendencies to disperse interactions across genders, implying our gender bias is unimportant and hence gender information will be used in the *demographics* feature vector.

Birthday breakdown:

Year	Frequency
Undisclosed	1
1901-1905	1
1906-1910	0
1911-1915	1
1916-1920	0
1921-1925	0
1926-1930	0
1931-1935	0
1936-1940	1
1941-1945	0
1946-1950	0
1951-1955	0
1956-1960	2
1961-1965	1
1966-1970	4
1971-1975	10
1976-1980	12
1981-1985	25
1986-1990	34
1991-1995	25
1996-2000	2

Table 4.2: Birthday breakdown for app users. There is a clear densely populated age range of around $18 - 30$.

Birthdays are grouped in a distinct range, most users in this data set are grouped in the range of around $18 - 30$. [Ugander and Marlow 2011] have found that there is a strong effect of age on friendship preferences, which implies that many of our app users should share similar preferences and hence birthday information will be a useful component of this feature.

Location breakdown:

Location	**Frequency**
Undisclosed	33
Ahmedabad, India	1
Bangi, Malaysia	1
Bathurst, New South Wales	1
Bellevue, Washington	1
Braddon, Australian Capital Territory, Australia	1
Brisbane, Queensland, Australia	2
Canberra, Australian Capital Territory	56
Culver City, California	1
Frederick, Maryland	3
Geelong, Victoria	1

Table 4.3: Location breakdown for app users. Most users are either undisclosed of based in Canberra where this app was developed.

Given the fact that most app users are either situated in Canberra (location of the app development and deployment) or are undisclosed, location information will not be used by this feature vector.

For *demographics* each user u, item v and feature vector x the alters set is defined as $alters_{u,v} = \{z|Liked_{z,v}\}$ the relationships R are explicitly defined below:

- $x_1 =$ Whether the user u is male, $gender_u = male$.
- $x_2 =$ Whether the user u is female, $gender_u = female$.
- $x_3 =$ Whether the user u and any user in the alters set share the same gender, $gender_u \in \{gender_{\{alters\}}\}$.
- $x_4 =$ Whether the user u and any user in the alters are of a different gender, $gender_u \notin \{gender_{\{alters\}}\}$.
- $x_5 =$ Whether the user u and any user in the alters set share the same birth range, $birthday_u \in \{birthday_{\{alters\}}\}$.

Where the functions $gender_u \in \{male, female\}$ returns the gender of u and $birthday_u \in range(1935 - 2000)$ returns the birthday range for u.

Applying this feature to our classifiers we obtain:

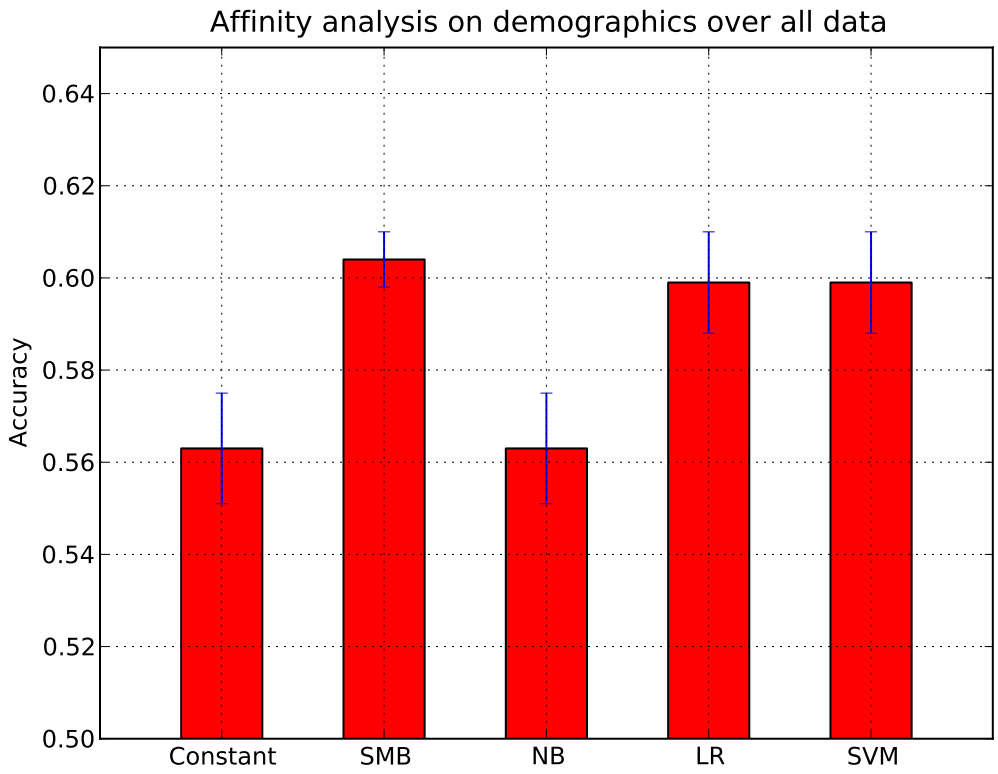

Figure 4.1: Accuracy results using *demographics* features. *Demographics* are our best performing feature so far for $k = 0$, however we still do not outperform our SMB baseline.

The *demographics* feature vector provides our best results so far for the case when $k = 0$, however it is still less predictive then our SMB baseline.

Comparing *demographics* against exposure we obtain:

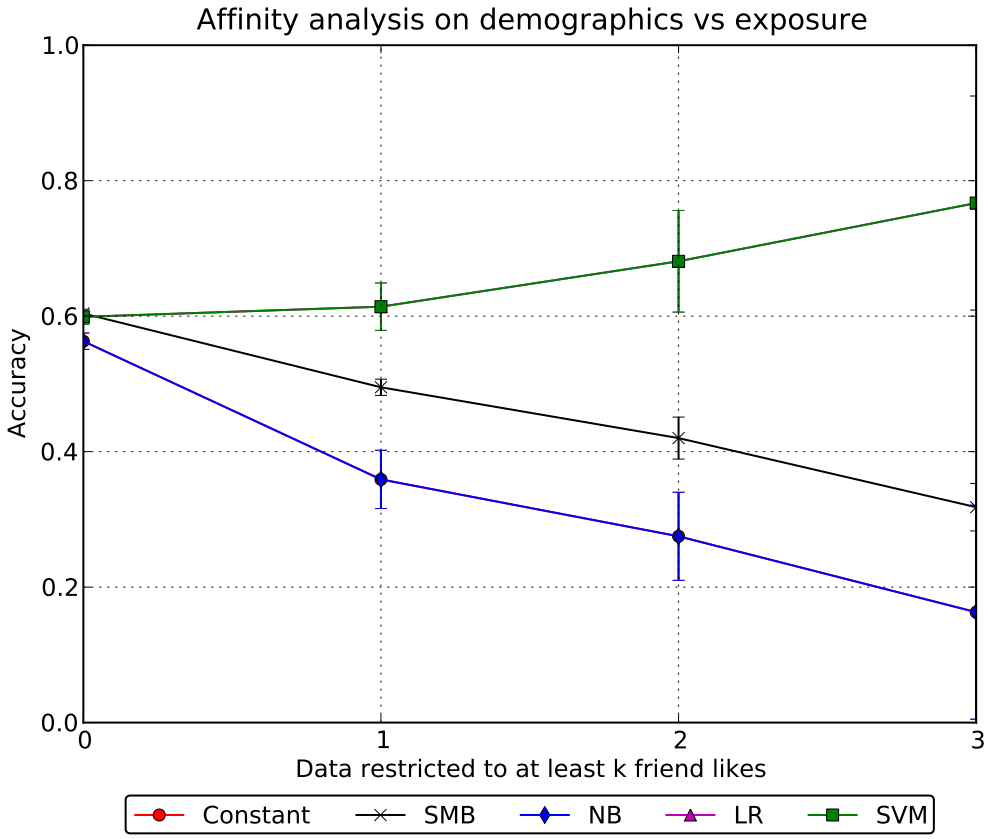

Figure 4.2: Accuracy results for exposure using the *demographics* features. As k increases the predictiveness of *demographics* substantially increases with it. Clearly gender and age are both predictive. Note in this case Constant = NB and LR = SVM.

The exposure for *demographics* shows a sizable improvement over our baselines as our k increases. This demonstrates that as the number of friends who like an item increases, the predictiveness of our *demographics* feature increases. Clearly, both gender and age are predictive *user features*.

4.2 Favourites

Facebook facilitates a wide variety of user selected *favourites*.

The *favourites* we will investigate include:

- **Activities**
- **Athletes**
- **Books**

- **Interests**
- **Movies**
- **Music**
- **People**
- **Sports**
- **Teams**
- **Television**

Memberships of different *favourite* categories can be partitioned into three sets (high, medium, low) based on membership frequency of our app users.

Below we display example tables of the the three different frequency categories: (A full table set can be found in *Appendix A*).

F	Television
20	The Big Bang [..]
19	How I Met [..]
14	The Simpsons
13	Top Gear
12	Futurama
12	Scrubs
11	Black Books
10	Black Books
10	South Park
10	Family Guy
9	The Daily Show
8	The IT Crowd
8	FRIENDS
7	True Blood
7	MythBusters

Table 4.4: Top *Television* shows for app users.

F	Interest
5	Movies
5	Music
3	Cooking
3	Sports
2	Psychology
2	Internet
2	Video Games
2	Martial arts
2	Literature
2	Economics
2	Tennis
2	Badminton
2	Artificial intelligence
2	Computers
2	Travel

Table 4.5: Top *Interests* for app users.

F	People
2	Alan Turing
1	Bender
1	Maurice Moss
1	Steve Jobs
1	Sean Parker
1	Pope Benedict XVI
1	Martin Luther
1	Alistair McGrath
1	St Augustine
1	Dennis Ritchie
1	Linus Torvalds
1	Richard Stallman
1	C. S. Lewis
1	Mike Oldfield
1	Ryan Giggs

Table 4.6: *Inspirational people* for app users.

The different frequency levels (high, medium, low) between *favourites* is summarised below:

- **High Locality**: *Music*, *Movies*, *Television* - Showing our app users appear to share similar *favourites* in a media setting.
- **Medium Locality**: *Activities*, *Books*, *Interests*, *Sports* - Showing our app users share some degree of similar preferences across these *favourites*.

- **Low Locality**: *Inspirational People*, *Athletes*, *Teams* - Showing our app users do not share many similar preferences across these *favourites* involving specific people and teams.

For *favourites* each user u and feature vector x is defined as:

$$I = \{Activities, Athletes, Books, Interests, Movies, Music, People, Sports, Teams, Television\}$$

The alters set for each i is conditioned by the relationship R, where:

$$R_{i,u,v} = \{z | Liked_{z,v} \wedge Favourite_{u,v,i}\}$$

In this case the *Favourite* function returns whether user u and user z both share a *favourite* via the current *favourite* i.

Applying this feature vector to our classifiers we obtain:

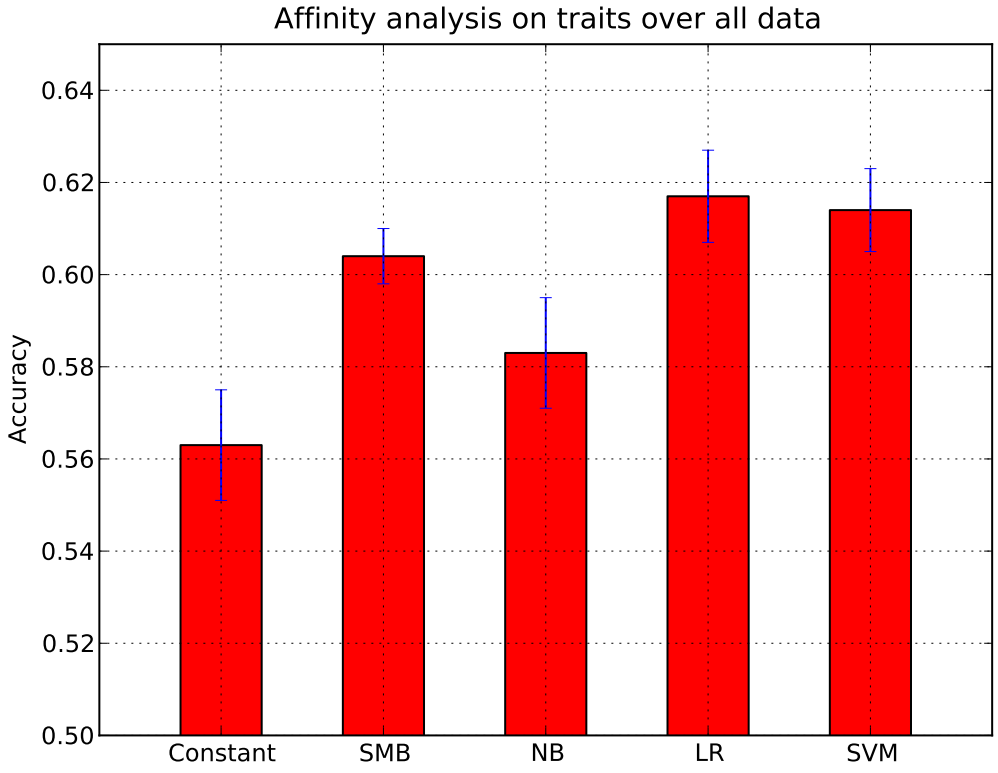

Figure 4.3: Accuracy results using the *favourites* feature. *Favourites* are our first feature more predictive then SMB for the case when $k = 0$, indicating that *favourites* are highly predictive of user preferences.

The *favourites* feature shows our first improvement over our SMB baseline in both the LR and SVM case for $k = 0$ demonstrating that *favourites* are more predictive then any previously applied feature.

Comparing *favourites* across our exposure we obtain:

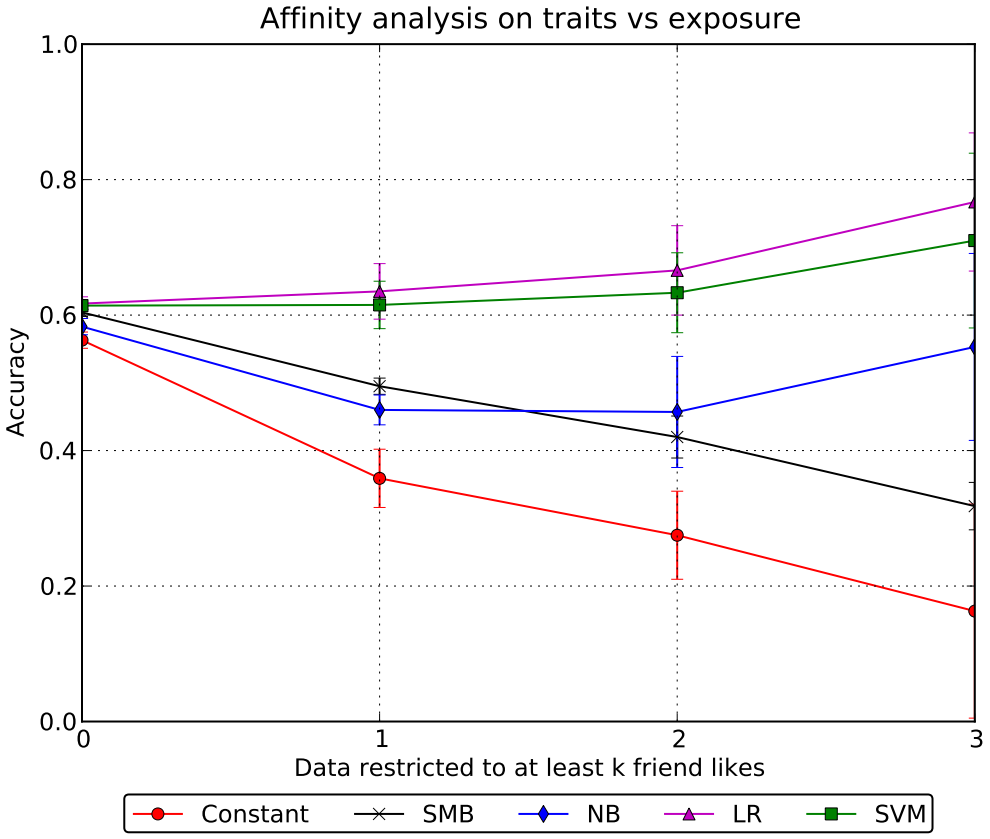

Figure 4.4: Accuracy results against exposure using the *favourites* features. Clearly, *favourites* are more predictive of user preferences compared with our baselines across all exposures k.

This predictive trend continues with exposure where each successive increase of k causes the performance of our classifiers to increase using the *favourites* features.

Given this predictive result, we extract the model weights for the most predictive classifier of LR from the case where $k = 0$. In the following table we can see which *favourites* contribute the most predictive qualities:

Favourite	**Weight**	**Frequency**
Activities	-5.927 $\pm$ 0.001	281
Television	-5.210 $\pm$ 0.0	1,029
Music	-3.409 $\pm$ 0.001	629
Movies	-2.668 $\pm$ 0.001	454
Interests	-1.921 $\pm$ 0.001	64
Sports	-1.820 $\pm$ 0.001	27
Books	-1.769 $\pm$ 0.0	163

Table 4.7: LR feature weights extracted for the case where $k = 0$. The *favourite* column displays the current *favourite*. The *weight* column shows the weighting given for this *favourite* and the *frequency* column displays the number of times this *favourite* was set to 1. We find that both high and medium frequency *favourites* are most predictive.

This table shows that the *favourites* which exhibit a high to medium frequency have a larger influence during prediction.

[Brandtzg and Nov 2011] found that virtual interactions help reveal common interests, while real world interactions help support friendships. Our data supports this, these common interests or *favourites* investigated above are clearly predictive of user likes, providing our most predictive results so far.

4.3 Groups

Facebook facilitates users to join *groups* for a large number of different types ranging from local sports teams and political preferences to computer games. These *groups* facilitate users to associate themselves with other who users who also share this same *group* preference.

The most popular groups for our app users are shown below:

The most popular *groups* joined by our app users exhibit a high degree of ANU and Canberra based locality. Additionally the frequencies of these top *groups* are consistent with the most predictive *favourites* outlined in the previous section.

Given the quantity of *groups* on Facebook, we need to find some optimal *group* test size j for our data set. Given memory and time constraints we tested within a range of $\{100 - 1000\}$ with an incremental step size of 100 for each test.

Frequency	Group Name
27	ANU StalkerSpace
20	Facebook Developers
15	ANU CSSA
14	CSSA
13	Australian National University
11	ANU - ML and AI Stanford Course
10	iDiscount ANU
10	Our Hero: Clem Baker-Finch
9	Students In Canberra
7	I grew up in Australia in the 90s
7	Grow up Australia - R18+ Rating for Computer Games
7	ANU Engineering Students' Association (ANUESA) 2010
7	ANU Postgraduate and Research Student Association (PARSA)
6	No, I Don't Care If I Die At 12AM, I Refuse To Pass On Your Chain Letter.
6	No Australian Internet Censorship
6	The Chaser Appreciation Society
6	Feed a Child with a Click
6	ANU Mathematics Society
6	ANU International Student Services, CRICOS Provider Number 00120C
6	2011 New & Returning Burton & Garran Hall
5	If You Can't Differentiate Between "Your" and "You're" You Deserve To Die
5	Keep the ANU Supermarket!!!
5	If 1m people join, girlfriend will let me turn our house into a pirate ship
5	The Great Australian Internet Blackout
5	When I was your age, Pluto was a planet.

Table 4.8: Popular *groups* breakdown for our app users. The *groups* joined by our app users exhibit a high degree of locality, many of these top rated *groups* have an ANU and Canberra focus.

The results for these tests are shown below:

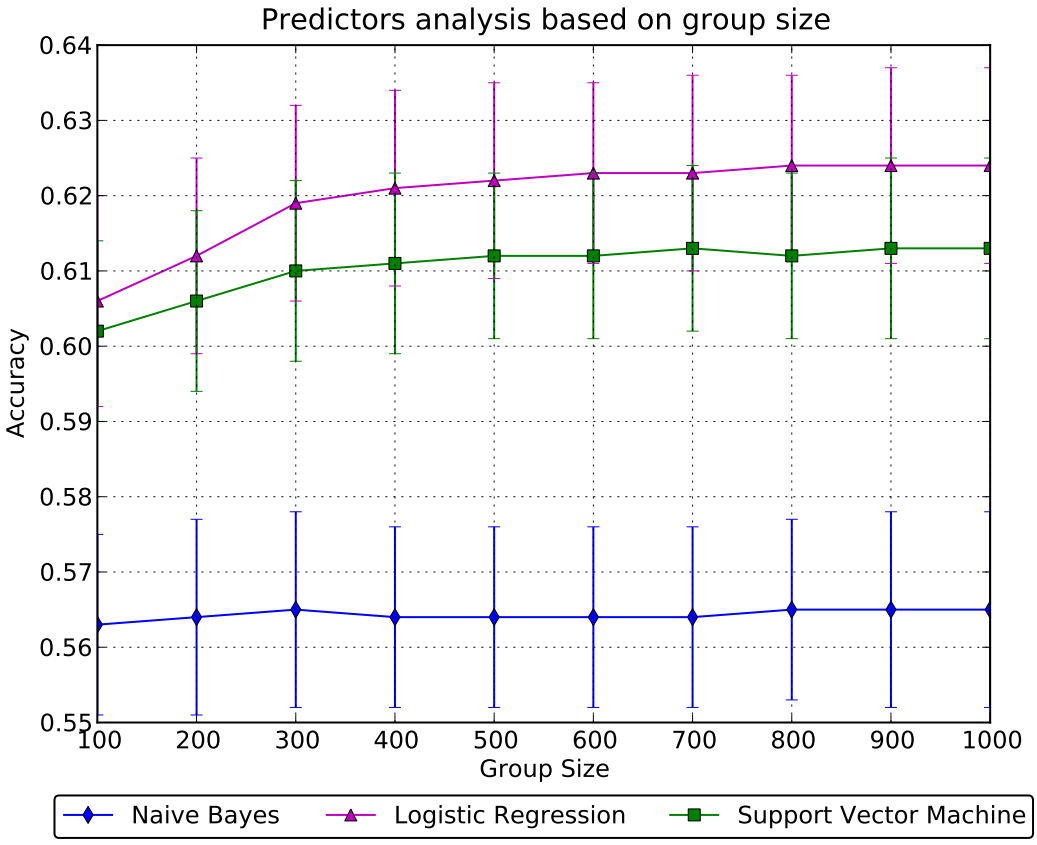

Figure 4.5: Accuracy results for testing using different *group* sizes. Unlike the previous results for the unpredictive *messages* features, the predictiveness of *groups* increases as the *groups* size increases, implying that the more *groups* the more predictive this feature will be.

Here we see as the *group* size increases, the predictiveness of this feature also improves. LR and SVM show a gradual increase as this *group* size increases, alluding to the possibility of an even higher group size being optimal.

The most predictive *group* sizes j for each of our classifiers are:

- **Naive Bayes**: 300
- **Logistic Regression**: 900
- **Support Vector Machine**: 800

For *groups* each user u and feature vector x is defined as:

$$I = \{Group\} \times \{Numberof Groups\}$$

Where the optimal *groups size* J is defined above for each classifier.

The alters set for each i is conditioned by the relationship R, where:

$$R_{i,u,v} = \{z | Liked_{z,v} \wedge GroupLiked_{i,u,z}\}$$

In this case the *GroupLiked* function returns whether both user u and user z have both liked the *group* at index i.

Using the most predictive *group* sizes j for each of our classifiers as defined above and comparing to our baselines we obtain:

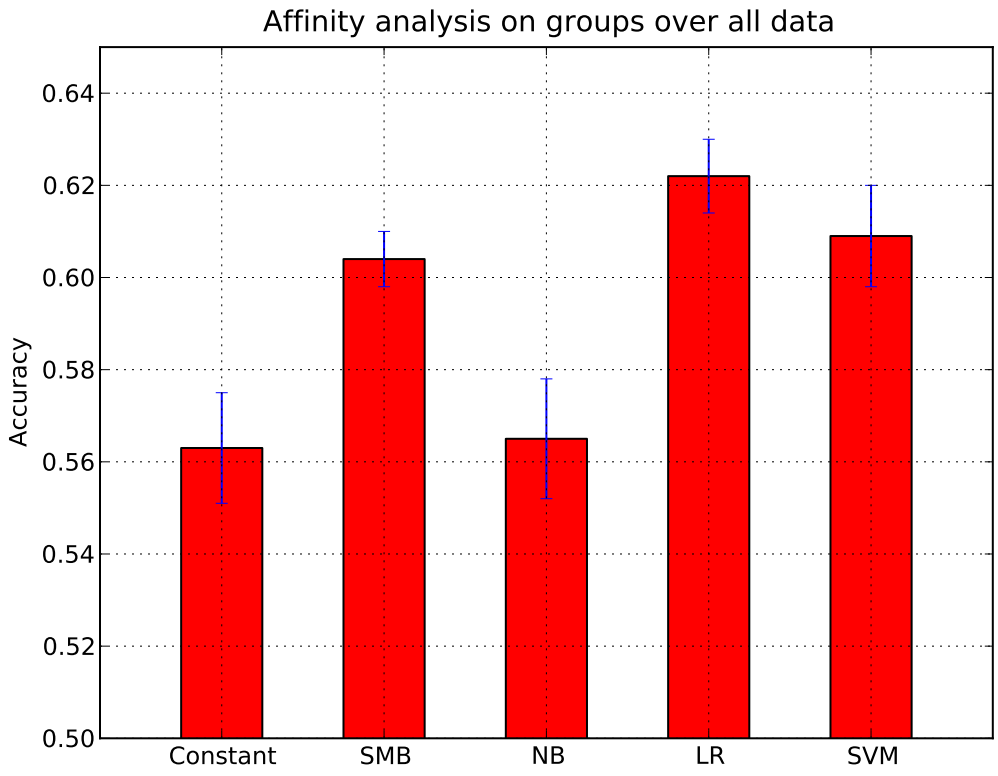

Figure 4.6: Accuracy results using the *groups* feature vector. We see *groups* are also more predictive over our baselines for the case where $k = 0$, particularly the LR classifier.

Both LR and SVM show an improvement over our SMB baseline for the case of $k = 0$ demonstrating that *groups* are predictive features.

Applying the *groups* feature against our exposure, we obtain:

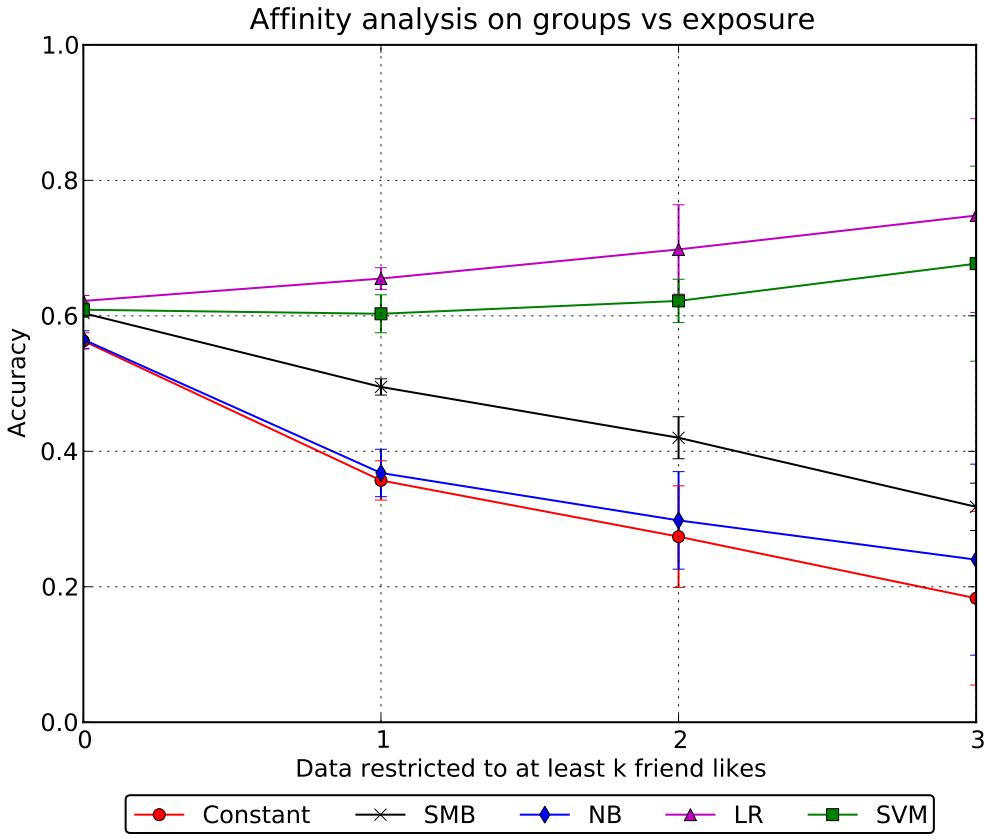

Figure 4.7: Accuracy results against exposure using the *groups* features. Here we see that the predictive trend for the $k = 0$ case continues with exposure for both LR and SVM.

This predictive trend continues across with exposure where each successive increase of k causes the performance of our LR and SVM classifiers to increase.

Given the predictive tendencies of groups, we extract the model weights for the case where $k = 0$ to see which *groups* contain the most predictive qualities:

Name	Size	Weight	Frequency
ANU StalkerSpace	1292	-7.236 $\pm$ 0	453
Facebook Developers	487	-3.442 $\pm$ 0	177
ANU CSSA	38	-2.742 $\pm$ 0	191
Australian National University	619	-2.565 $\pm$ 0	70
Overheard at the Ateneo de Manila University	253	-2.462 $\pm$ 0	26
iDiscount ANU	338	-2.203 $\pm$ 0	88
PETITION FOR FACEBOOK TO INSTALL A DISLIKE BUTTON	683	-2.018 $\pm$ 0	92
I grew up in Australia in the 90s	731	-1.991 $\pm$ 0	75
Grow up Australia - R18+ Rating for Computer Games	222	-1.951 $\pm$ 0	102
Heavy Metal - CANBERRA METAL	30	-1.694 $\pm$ 0	42

Table 4.9: LR feature weights extracted for the case where $k = 0$. The *name* column displays the current *group* name. The *size* column shows the total size of this group across all users. The *weight* column shows the weighting given for this *group* and the *frequency* column displays the number of times this *group* was set to 1. We find that highly local *groups* with a high frequency of app users are most predictive.

Groups are highly predictive of user preferences, we see from the weights table above that *groups* which are highly local to the ANU and Canberra are are highly predictive as well as *groups* which have a high frequency of app users.

4.4 Pages

Facebook facilitates users to like *pages* for 'things' they like across a wide spectrum of different areas ranging from web browsers and TV shows to schools. This allows Facebook users to associate themselves with other users who like these similar 'things'.

The most popular *pages* liked by our app users are shown below:

Frequency	**Page Name**
33	ANU Computer Science Students' Association (ANU CSSA) 2011
32	The Australian National University
31	ANU Stalkerspace
21	Humans vs Zombies @ ANU
20	The Big Bang Theory
19	Australian National University
19	How I Met Your Mother
18	ANU LinkR
18	ANU ducks
17	Australian National University Students' Association
16	Google
15	Google Chrome
15	ANU XSA
15	Facebook
14	YouTube
14	The Simpsons
13	Portal
13	Top Gear
13	Music
13	ANU Memes
12	Futurama
12	Scrubs
12	ANU O-Week 2012: Escape to the East
12	The Stig
11	Black Books

Table 4.10: Popular *pages* breakdown for our app users. *Pages* exhibit less locality preferences when compared with *pages*, while some *pages* are Canberra focused, many are also more general. Additionally app user memberships for *pages* is much higher then for *groups*.

In comparison with *groups* and *favourites*, *pages* show a frequency across the most popular *pages* for app users and have a less Canberra and ANU centric focus.

Given the quantity of *pages* on Facebook, we need to find some optimal test size j for each classifier. Given memory and time constraints we tested within a range of $(100 - 1000)$ with an incremental step size of 100 for each test.

The results of these tests are shown below:

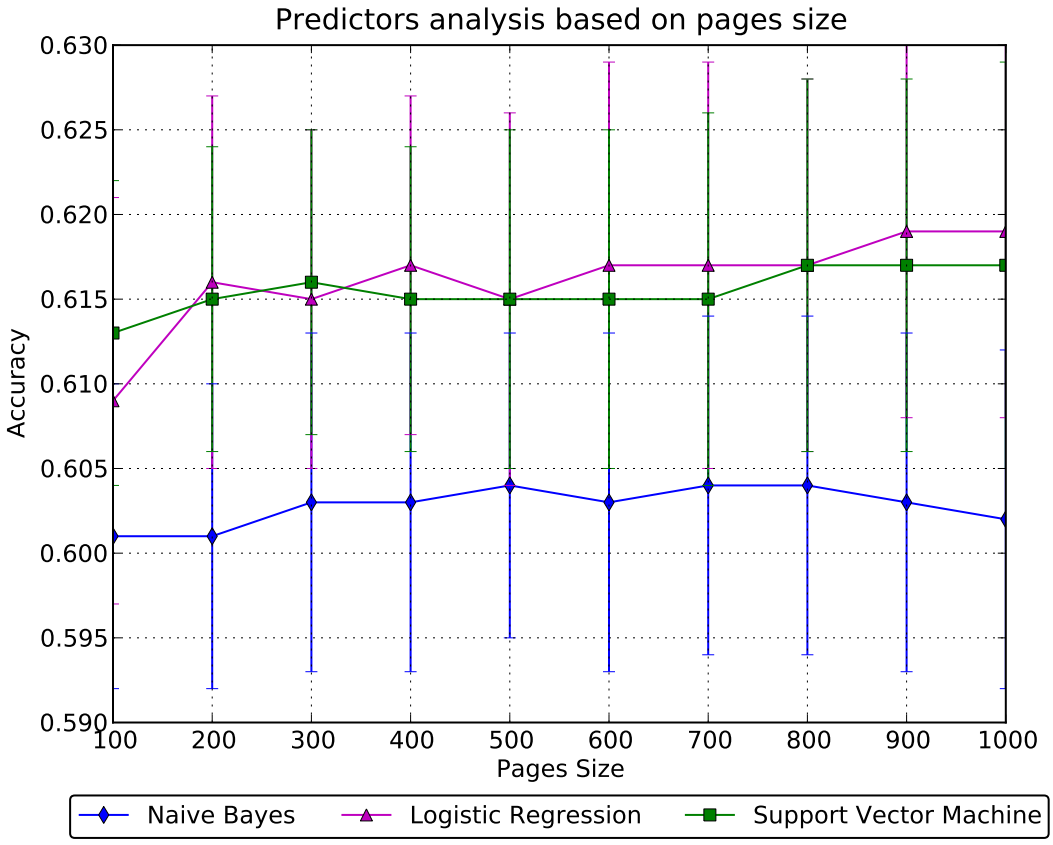

Figure 4.8: Accuracy results for different *page* sizes. Results for *page* sizes are quite jumpy, however a similar trend from *groups* follows that the more *pages* we use for testing, the more predictive our results.

The most predictive *page* sizes j for each of our classifiers are:

- **Naive Bayes**: 500
- **Logistic Regression**: 900
- **Support Vector Machine**: 800

LR and SVM exhibit a gradual increase as our *page* size increases, alluding to the possibility of an even higher *page* size being optimal for prediction.

For *pages* each user u and feature vector x is defined as:

$$I = \{Page\} \times \{Numberof Pages\}$$

Where the optimal *pages size* J is defined above for each classifier.

The alters set for each i is conditioned by the relationship R, where:

$$R_{i,u,v} = \{z | Liked_{z,v} \wedge PageLiked_{i,u,z}\}$$

In this case the *PageLiked* function returns whether both user u and user z have both liked the *page* at index i.

Using the most predictive *page* sizes j for each of our classifiers as defined above we obtain:

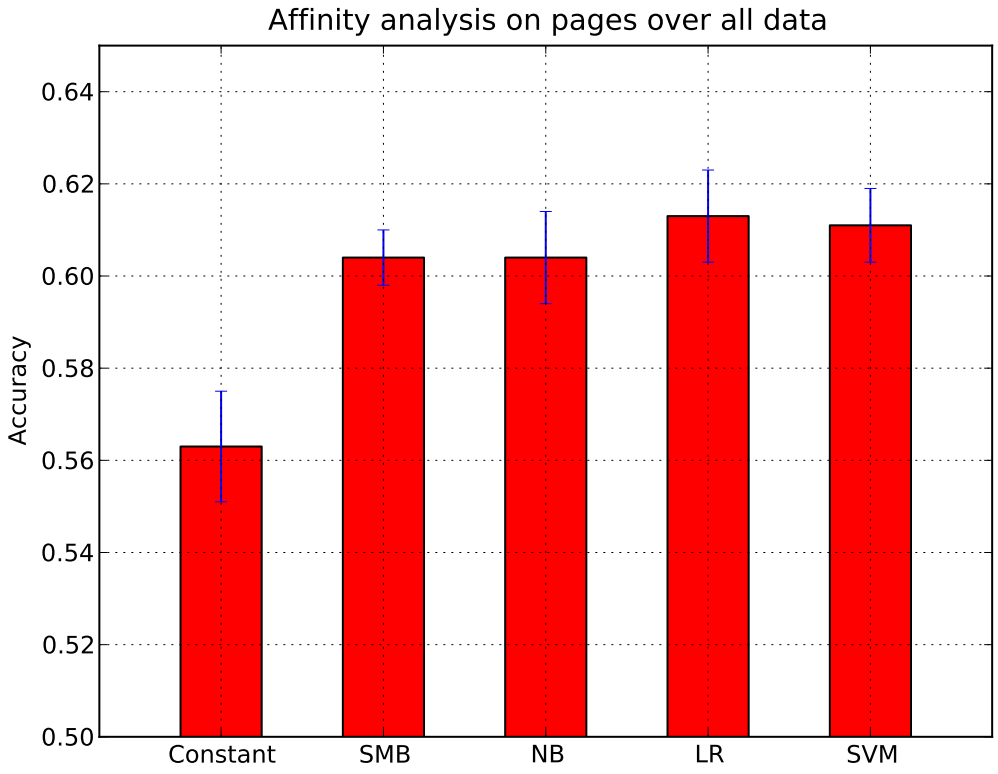

Figure 4.9: Accuracy results using the *pages* feature, we can see an improvement over our baselines for both LR and SVM. Demonstrating that *pages* are a predictive *user preference*, however they are not as predictive as *groups* or *favourites*.

Here we see NB, LR and SVM show an improvement over our SMB baseline for the case of $k = 0$ demonstrating that *pages* are also predictive. However, *groups* and *favourites* are still both more predictive. Possibly due to the fact that *groups* are generally more local and they have a lower frequency of app user members.

Applying the *pages* feature against exposure, we obtain:

Figure 4.10: Accuracy results against exposure using the *pages* feature. We see a similar trend as demonstrated in *groups* and *favourites* where the predictiveness of this feature improves with exposure k.

The trend of improved predictiveness over exposure k continues similarly with *pages*. As more data is available for the classifiers to learn from, their prediction accuracy increases.

By extracting the model weights from the case where $k = 0$ we can see which *pages* contain the most predictive qualities:

Name	**Size**	**Weight**	**Frequency**
Sorry mate i can't, i've got Quidditch	254	-1.799 $\pm$ 0	18
Avatar: The Last Airbender	324	-1.514 $\pm$ 0.001	13
National Geographic	662	-1.437 $\pm$ 0.001	18
The Simpsons	1552	-1.414 $\pm$ 0	170
Sushi	387	-1.33 $\pm$ 0.001	9
House	1746	-1.291 $\pm$ 0	66
Seinfeld	609	-1.249 $\pm$ 0	15
Starbucks	1548	-1.249 $\pm$ 0	7
American Dad	540	-1.215 $\pm$ 0.001	18
friends don't let friends vote for Tony Abbott	551	-1.206 $\pm$ 0.001	19

Table 4.11: Negative LR feature weights extracted for the case where $k = 0$. The *name* column displays the current *page* name. The *size* column shows the total size of this group across all users. The *weight* column shows the weighting given for this *page* and the *frequency* column displays the number of times this *page* was set to 1. We find non-local and medium sized *pages* to be most predictive.

Name	**Size**	**Weight**	**Frequency**
CatDog	259	1.815 $\pm$ 0.001	12
Worst. Idea. Ever. [pause] Let's do it.	227	1.737 $\pm$ 0	21
Grug	279	1.698 $\pm$ 0	9
Kings Of Leon	840	1.607 $\pm$ 0.001	14
Planking Australia	166	1.598 $\pm$ 0.001	4
Dr. House	964	1.588 $\pm$ 0	28
Suit Up	466	1.389 $\pm$ 0.001	17
Don't you hate it when Gandalf marks your [..]	110	1.372 $\pm$ 0.001	19
Paramore	1004	1.343 $\pm$ 0.001	31
Tintin	250	1.339 $\pm$ 0.001	11

Table 4.12: Positive LR feature weights extracted for the case where $k = 0$. The *name* column displays the current *page* name. The *size* column shows the total size of this group across all users. The *weight* column shows the weighting given for this *page* and the *frequency* column displays the number of times this *page* was set to 1. We find non-local and medium sized *pages* to be most predictive.

The most predictive *pages* in our data are non-local, which was the opposite case to *groups*, additionally medium sized *pages* with highly varied frequencies were the most predictive.

4.5 Conclusion

Throughout this chapter we have explored different *user preferences* available to users to demonstrate their personal preferences across a range of different topics and mediums.

Others have pointed out that non-social information is more predictive of user likes [www] and we observe that too. We have found that *user preferences* are more predictive of user likes compared with *user interactions*, this is particularly true for *favourites*, *groups* and *pages*. This observation holds true for the exposure case of $k = 0$ and our predictions continue to improve with each successive increase in k (notably at the detriment of our baseline).

Chapter 5

Feature Combination

Given the vast number of potential affinity features as outlined in the previous sections, it is both computationally costly and time consuming to group all features together into one huge combined affinity feature. Hence, it is crucial we provide a practical method which facilitates a combination of the most predictive individual affinity features found during this research.

In this chapter we combine the individually most predictive affinity features together into one large, but manageable feature vector and examine the results.

5.1 Affinity Feature Selection

Based on results found during our *user preference* analysis, the combined affinity feature we examine is comprised of:

- *Favourites*
- *Groups*
- *Pages*

Applying this combined affinity feature to the data set:

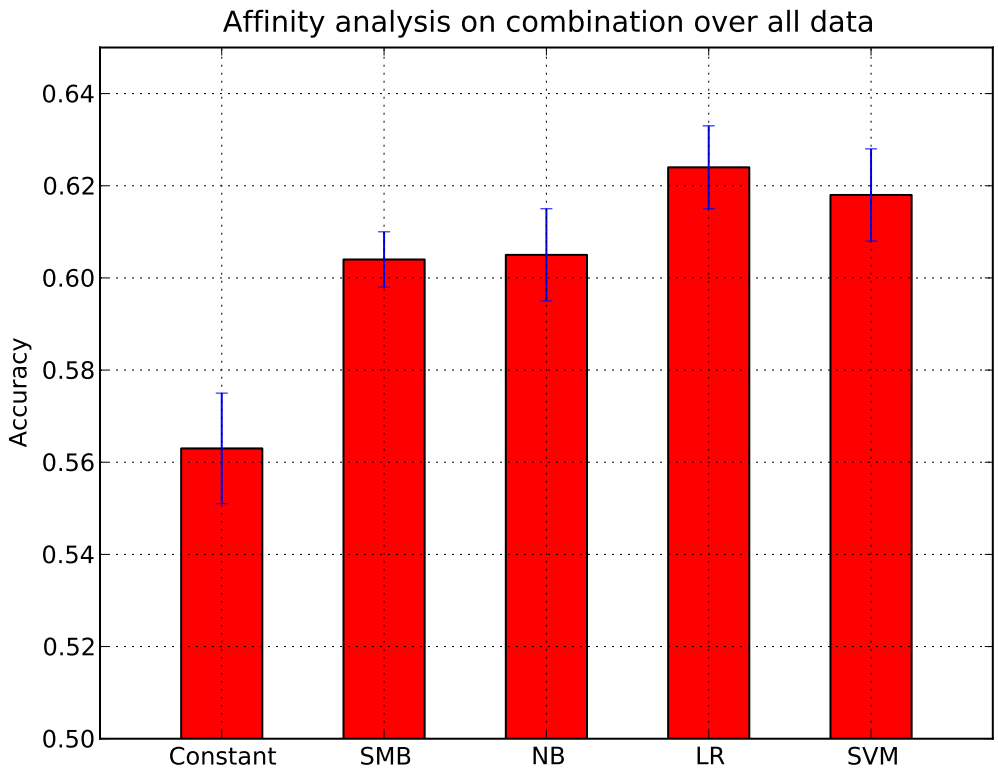

Figure 5.1: Accuracy results using the *combined* feature set. We can see a large improvement over our SMB baseline for the case where $k = 0$, particularly for the LR case.

We find that the *combined* affinity feature gives better results for our classifiers when compared with our baselines, particularly for the LR case. In fact, this *combined* affinity feature results in the best results found during this research, which is summarised in the table below:

Classifier	**Accuracy**
NB	0.583 ± 0.012
LR	0.617 ± 0.01
SVM	0.614 ± 0.009

Table 5.1: *Favourite* feature results for $k = 0$.

Classifier	**Accuracy**
NB	0.604 ± 0.01
LR	0.613 ± 0.01
SVM	0.611 ± 0.008

Table 5.2: *Pages* feature results for $k = 0$.

Classifier	**Accuracy**
NB	0.565 ± 0.013
LR	0.622 ± 0.008
SVM	0.609 ± 0.011

Table 5.3: *Groups* feature results for $k = 0$.

Classifier	**Accuracy**
NB	0.605 ± 0.01
LR	0.624 ± 0.009
SVM	0.618 ± 0.01

Table 5.4: *Combined* feature results for $k = 0$.

These results show that based on our results the most predictive feature vector is a combination of the individual most predictive feature vectors found in our *user preferences* section.

This trend continues for all values of k and offers the most predictive feature vector found during this research.

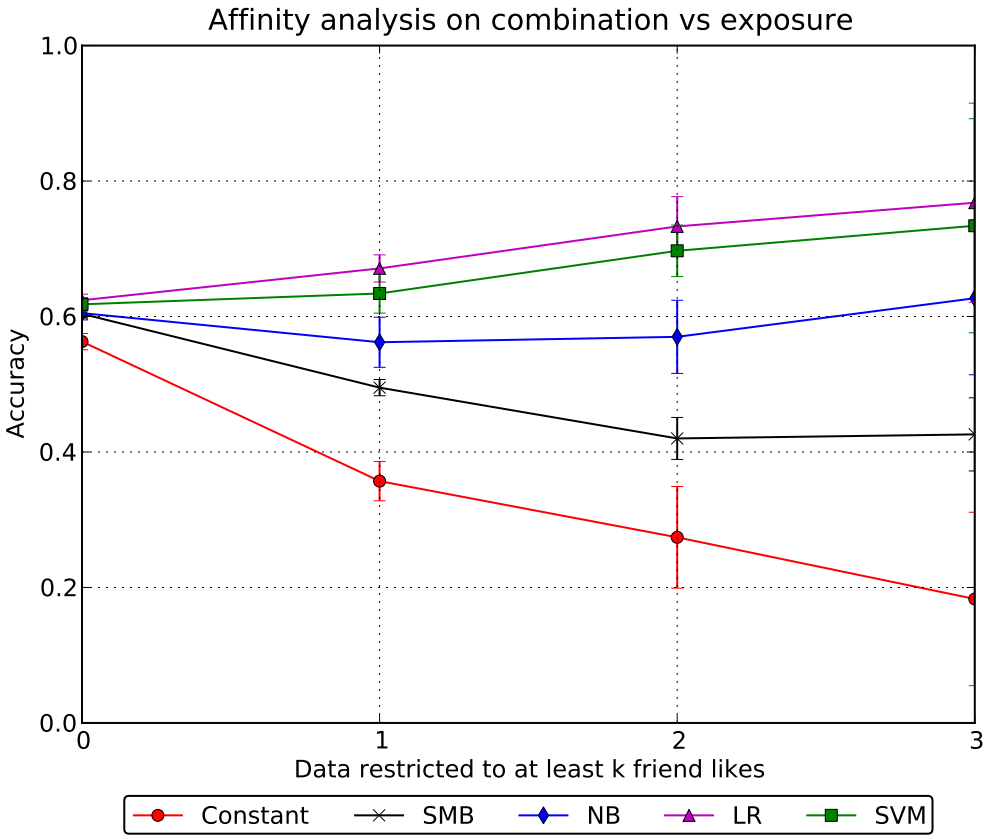

Figure 5.2: Accuracy results across exposure using the *combined* feature set. This feature set offers the most predictive results discovered during this research. As our exposure k increases, the improved predictive performance of this feature combination increases with it.

Clearly, this *combined* feature provides the most predictive results when compared with all other analysis completed during this thesis.

By extracting the model feature weights from the case where $k = 0$ we can see which components of the *combined* affinity feature were most predictive:

Name	**Size**	**Weight**	**Frequency**
Avatar: The Last Airbender (Page)	324	-1.68 $\pm$ 0.001	13
I'm late. Got attacked by a wild Pokemon (Page)	161	-1.609 $\pm$ 0	20
Overheard at the Ateneo de Manila (Group)	253	-1.527 $\pm$ 0.001	26
Sorry mate i can't, i've got Quidditch (Page)	254	-1.501 $\pm$ 0	18
I would.....for Escapium. (Group)	50	-1.467 $\pm$ 0.001	11
Burgtoons (Group)	34	-1.37 $\pm$ 0.001	7
The Simpsons (Page)	1552	-1.355 $\pm$ 0.001	170
City Gate Hall (Group)	27	-1.346 $\pm$ 0	5
Victoria's Secret (Page)	764	-1.337 $\pm$ 0	11
Starbucks (Page)	1548	-1.313 $\pm$ 0	7

Table 5.5: *Logistic Regression* feature weights extracted for the negative case where $k = 0$. The *Name* column displays the name of the feature. *Size* represents the size of the *Page* or *Group*. *Weight* represents the weight this feature vector received. *Frequency* displays the number of times this feature vector was set to 1 for a user.

Name	**Size**	**Weight**	**Frequency**
Don't you hate it when Gandalf marks [...] (Page)	110	1.627 $\pm$ 0.001	19
Goodberry's (Page)	318	1.591 $\pm$ 0	73
Worst. Idea. Ever. [pause] Let's do it. (Page)	227	1.561 $\pm$ 0	21
CatDog (Page)	259	1.531 $\pm$ 0.001	12
Planking Australia (Page)	166	1.501 $\pm$ 0.001	4
Avenged Sevenfold (Page)	351	1.471 $\pm$ 0	6
Grug (Page)	279	1.465 $\pm$ 0	9
Dr. House (Page)	964	1.451 $\pm$ 0	28
If 1m people join, girlfriend will let [...] (Group)	416	1.362 $\pm$ 0	68
Do you ride kangaroos? no mate the [...] (Page)	321	1.333 $\pm$ 0.001	23

Table 5.6: *Logistic Regression* feature weights extracted for the positive case where $k = 0$. The *Name* column displays the name of the feature. *Size* represents the size of the *Page* or *Group*. *Weight* represents the weight this feature vector received. *Frequency* displays the number of times this feature vector was set to 1 for a user.

The negative LR weights are equally broken up into *pages* and *groups* as contributing highly to the prediction. The size ranges for both *pages* and *groups* varies greatly from as little as 27 up to as high as 1552 and there is little corelation between the frequency and sizes, in fact lower frequencies 20 contribute most to the prediction.

The positive LR weights are more focused on the predictiveness of *pages* and the sizes of the *pages* are much more consistent 200.

These LR weights show that both *groups* and especially *pages* are highly predictive of a users like preferences and particular sizes or frequencies do not appear to be more predictive then others. Additionally under this combined paradigm, *favourites* do not

appear to contribute strongly to our results.

We have shown in this chapter that a *combined* version of our individually predictive affinity features results in the most accurate predictions for our data set, when compared with all other individual features.

Chapter 6

Conclusion

In this chapter we will outline a summary of the work completed during this thesis and outlines a proposal for future work in this area.

method improves with exposure

6.1 Summary

In this thesis we have tested and compared an exhaustive list of different affinity features across varied exposures sizes.

We have demonstrated that *user interaction* affinity features in themselves are not predictive of user likes, however coupled with user exposure, they show a comprehensive improvement over our baselines.

We have also shown the interesting result that *user preference* affinity features are more predictive of user likes when compared with our baselines and this trend continues with user exposure.

To answer the question initially proposed for this thesis, we have shown the affinity features which provide the highest predictiveness of user likes come from *user preferences* and not *user interactions*. The most predictive features found in this analysis are *favourites*, *group memberships* and *page likes*.

This provides the exciting and novel insights examined during this thesis.

6.2 Future Work

Proposed future work can be summarised under the following points:

- **Increase size ranges**: Given our maximum test sizes for *groups* and *pages* of 1000 this size could be increased to find the optimal testing range for each of our classifiers.

- **Passive likes**: Given the Facebook model of allowing users to like but not dislike data, explicit dislike data can not be gleaned from Facebook, which is why NICTA sourced active likes data was used for this evaluation. An approach could be developed which can predict whether a user will have seen an item

(online timestamps, recent interactions with user) and can infer that if the user did not like the item then they disliked it. This data set could then be applied to the testing methodology undertaken above.

- **Cold start**: Leaving out some subset of users when training our models, but including them during testing to explore their effects on results.
- **General user set**: Such as the study done by [Ugander and Marlow 2011] which comprised of the entire active social network of 721 million users (as of May 2011), applying these methods to a data set which is more indicative of the general Facebook user population could offer more generalisable results.
- **Bayesian Model Averaging**: Weighting the most successful machine learning models under different affinity features and exposures to generate a new combined classifier, which combines the best results of each individual classifier.

Appendix A

Favourites Group Summary

Frequency	Activity
10	Sleeping
5	Eating
5	Reading
4	Running
4	Cycling
4	Minecraft
4	Programming
3	Android
3	Cooking
3	Video Games
3	Xbox 360
3	Piano
3	Guitar
3	Badminton
3	Chocolate

Table A.1: Top *Activities* for app users.

Frequency	Inspirational People
2	Alan Turing
1	Bender
1	Maurice Moss
1	Steve Jobs
1	Sean Parker
1	Pope Benedict XVI
1	Martin Luther
1	Alistair McGrath
1	St Augustine
1	Dennis Ritchie
1	Linus Torvalds
1	Richard Stallman
1	C. S. Lewis
1	Mike Oldfield
1	Ryan Giggs

Table A.2: Top *Inspirational People* for app users.

Frequency	Book
7	Harry Potter
4	The Bible
3	Harry Potter series
3	Discworld
3	That's 3 minutes of solid study, think I've earned 2hrs of Faceboook time
3	Freakonomics
3	Tomorrow when the War Began
2	Magician
2	Hitchhiker's Guide To The Galaxy
2	The Discworld Series
2	Terry Pratchett
2	Terry Pratchett
2	George Orwell
2	Lord Of The Rings
2	Goosebumps

Table A.3: Top *Books* for app users, here we see an example of the non-distinct properties inherent in Facebook, where books can have the same name, yet still be regarded as a different entity.

Frequency	Interest
5	Movies
5	Music
3	Cooking
3	Sports
2	Psychology
2	Internet
2	Video Games
2	Martial arts
2	Literature
2	Economics
2	Tennis
2	Badminton
2	Artificial intelligence
2	Computers
2	Travel

Table A.4: Top *Interests* for app users.

Frequency	Music
9	Daft Punk
9	Muse
8	Michael Jackson
8	Pink Floyd
8	Lady Gaga
7	Linkin Park
7	Avril Lavigne
6	Radiohead
6	Rihanna
6	Coldplay
6	Green Day
6	Katy Perry
6	Taylor Swift
5	Gorillaz
5	Queen

Table A.5: Top *Music* for app users.

Frequency	Movie
9	Inception
8	Avatar
8	Fight Club
7	The Lord of the Rings Trilogy (Official Page)
6	Star Wars
6	I wouldnt steal a car, But i'd download one if i could
6	WALL-E
6	Scott Pilgrim vs. the World
6	Toy Story
6	Shrek
5	Batman: The Dark Knight
5	Harry Potter
4	The Matrix
4	The Social Network Movie
4	Monsters, Inc.

Table A.6: Top *Movies* for app users.

Frequency	Sport
8	Badminton
5	Basketball
3	Cycling
3	Volleyball
2	Starcraft II
2	Football en salle
2	Swimming
2	Towel Baseball
2	Tennis
1	Soccer
1	Taekwondo
1	Rock climbing
1	In The Groove
1	Darts
1	Table tennis

Table A.7: Top *Sports* for app users.

Frequency	Television Show
20	The Big Bang Theory
19	How I Met Your Mother
14	The Simpsons
13	Top Gear
12	Futurama
12	Scrubs
11	Black Books
10	Black Books
10	South Park
10	Family Guy
9	The Daily Show
8	The IT Crowd
8	FRIENDS (TV Show)
7	True Blood
7	MythBusters

Table A.8: Top *Television* shows for app users.

Frequency	Athlete
4	Roger Federer
4	Rafael Nadal
3	Maria Sharapova
2	Leo Messi
1	Andy Schleck
1	Chrissie Wellington
1	Emma Snowsill
1	Emma Moffatt
1	Barbara Riveros
1	The Brownlee Brothers
1	Marie Slamtoinette #1792
1	Wayne Rooney
1	”you are what you eat” ” I dont remember eating a Tank.”
1	Nemanja Vidic
1	Ryan Giggs

Table A.9: Top *Athletes* for app users.

Frequency	Team
5	Manchester United
2	Bear Grylls cameraman appreciation society
2	Real Madrid C.F.
2	Liverpool FC
1	Leopard Trek
1	British Triathlon
1	TeamCWUK
1	Surly Griffins
1	Canberra Raiders
1	Kolkata Knight Riders
1	Brisbane Roar FC
1	Brisbane Broncos
1	Cricket Australia
1	— Manchester United Fans —
1	Juventus

Table A.10: Top *Teams* for app users.

Bibliography

New objective functions for social collaborative filtering. (pp. 13, 22, 42)

ALIAS-I. 2008. LINGPIPE 4.1.0. HTTP://ALIAS-I.COM/LINGPIPE (ACCESSED OCTOBER 1, . 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 27:1–27:27. (p. 10)

ANDERSON, A., HUTTENLOCHER, D., KLEINBERG, J., AND LESKOVEC, J. 2012. Effects of User Similarity in Social Media. In *ACM International Conference on Web Search and Data Mining (WSDM)* (2012). (p. 22)

BACKSTROM, L., BAKSHY, E., KLEINBERG, J., LENTO, T., AND ROSENN, I. 2011. Center of attention: How facebook users allocate attention across friends. ICWSM'11 (2011). (p. 23)

BRANDTZG, P. B. AND NOV, O. 2011. Facebook use and social capital — a longitudinal study. ICWSM'11 (2011). (p. 31)

CHANG, C.-C. AND LIN, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 27:1–27:27. (p. 10)

CUI, P., WANG, F., LIU, S., OU, M., AND YANG, S. 2011. Who should share what? item-level social influence prediction for users and posts ranking. In *International ACM SIGIR Conference (SIGIR)* (2011). (p. 9)

GRANOVETTER, M. S. 1978. Threshold models of collective behavior. *Am. J. Sociol 83(6):14201443*. (p. 2)

HILL, R. AND DUNBAR, R. 2003. Social network size in humans. *Human Nature 14*, 1, 53–72. (p. 1)

LANG, K. 1995. NewsWeeder: Learning to filter netnews. In *12th International Conference on Machine Learning ICML-95* (1995), pp. 331–339. (p. 8)

NOEL, J. G. 2011. New social collaborative filtering algorithms for recommendation on facebook (2011). (pp. 8, 9)

PANTEL, A., GAMON AND HAAS. 2012. *Proceeding SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. (p. 2)

RESNICK, P. AND VARIAN, H. R. 1997. Recommender systems. *Communications of the ACM 40*, 56–58. (p. 8)

SAEZ-TRUMPER, D., NETTLETON, D., AND BAEZA-YATES, R. 2011. High correlation between incoming and outgoing activity: A distinctive property of online

social networks? In *Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM '11 (2011). (p. 13)

SANGHVI, R. AND STEINBERG, A. 2010. Edgerank: The secret sauce that makes facebook's news feed tick (2010). (p. 1)

UGANDER, B., KARRER AND MARLOW. 2011. The anatomy of the facebook social graph. *CoRR abs/1111.4503*. (pp. 23, 24, 50)

WATTS, D. J. AND STROGATZ, S. H. 1998. Collective dynamics of small-world networks. *Nature 393(6684):440442*. (p. 2)

YANG, LONG, SMOLA, SADAGOPAN, ZHENG, AND ZHA. 2011. Like like alike: Joint friendship and interest propagation in social networks. In *WWW-11* (2011). (p. 9)