Interaction 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 2_ε.

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

Riley Kidd
13 October 2012

Abstract

Social networks provide a wide array of user specific interactions, profile information and user preferences. This thesis attempts to decipher which user interactions or preferences are truly indicative of 'likes', this information is then leveraged to allow for binary classification of user specific links with the goal of discovering the ideal combination of traits for prediction.

The success of our predictions are evaluated using a number of machine learning algorithms including, *Naive Bayes, Logistic Regression and Support Vector Machines*, results are compared to previous work using *Matchboxing and Social Matchboxing* techniques as baselines. The data set is sourced from a set of over 100 Facebook users and their interactions with over 30,000 friends during a four month period.

Our analysis has shown that user interactions in themselves are not predictive of user likes, however user preferences are. These results, coupled with user like exposure curves offer a useful paradigm for extracting and exploiting user preferences for prediction across our data set.

Contents

Chapter 1

Introduction

An individuals 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 social interactions (messages, posting, commenting, etc) and preferences (demographics, group memberships, likes, etc) is an ever expanding graph structure of verbose social content. The internet is becoming a network of people, providing a myriad of expanding social information and user driven content.

The ultimate question we wish to address in this thesis then becomes: How can we exploit this information to decipher which user interactions or preferences are most indicitive of user likes?

We address this question by comparing and contrasting different feature vectors in our data against appropriate baselines and ultimately offer a feature vector combination paradigm which represents improved results which are computationaly faster and can more appropriately generalise over a population then previous methods.

1.1 Objectives

The primary objective of this thesis is to contrast and compare differing user feature vectors across user interactions and preferences. Using the machine learning concepts of *Naive Bayes* (NB), *Logistic Regression* (LR) and *Support Vector Machines* (SVM) compared with our appropriate baslines of *Social Mathbox* (SMB) and *Constant Classifiers*. With the goal of discovering which features are most predictive or user likes.

Based on the insight that social inuence can play a crucial role in a range of behavioral phenomena [Granovetter 1978; Watts and Strogatz 1998] we will also test using an exposure curve [Romero et al. 2011] hold out technique, where data is only tested if some friend has already liked that datum.

Finally, we will analyse the effect of combining successful user feature vectors together using a *Feature Combination* approach.

1.2 Contributions

Our specific contributions made during this thesis show:

- Both *Interactions* and *Incoming*
  Outgoing messages are not more predictive then previously used *Social Matchboxing* techniques.
- Each user preference of *Traits*, *Groups* and *Pages* contributed to a better result then previously used techniques.
- Combining user preferences with an exposure curve for user likes results in a substantial improvement from previously used techniques.
- Combination of beneficial user preferences results in the most advantageous feature vector for this data.

Overall, we provide a methodology which improves upon previous work and offers an approach to combine positively contributing aspects of different feature vectors in our data.

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 feature vectors for user interactions and the results of applying these feature vectors in comparison with our baselines.
- **Chapter 4**: A similar feature vector analysis as above, however the feature vectors we utilise are for user preferences.
- **Chapter 5**: In this chapter we discuss results from combining different positive feature vectors and propose an ideal combination based on our analysis.
- **Chapter 6**: Finally, we draw the work done throughout this thesis to a conclusion and offer avenues for future work in this area.

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

Chapter 2

Background

In the following, we define the source of our data set, notation used throughout this thesis, our choice of prediction algorithms and our testing appraoch and methodology.

2.1 Facebook

Facebook is the largest and most active social media service in the world. Facebook users can create a profile containing personal preferences and have friendships and interactions between other users. These interactions can also be liked or commented on by other users.

discuss feed time for facebook

need to show all different ways to interact

2.2 LinkR

NICTA developed a Facebook app named LinkR [1] which would make recommendations to app users and record whether or not the user liked the item.

The dataset includes information about each app user as well as a subset of available data about their friends.

The LinkR Facebook app was used to collect information about users, their interactions and preferences. The data set contains information about app users as well as a sub-set of visible information about their friends. The app tracked and stored information for over 100 app users and their 39,000+ friends.

The four main interactions between users are posts (posting an element 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 table below outlines data collected during app trials.

The table below summarises the data collected from both app users and their friends.

Pertanent user data which is collected by the LinkR app includes:

[1] The main developer of the LinkR Facebook App is Khoi-Nguyen Tran, a PhD student at the Australian National University. Khoi-Nguyen wrote the user interface and database crawling code for LinkR.

App Users	**Posts**	**Tags**	**Comments**	**Likes**
Wall	27,955	5,256	15,121	11,033
Link	3,974	-	5,757	4,279
Photo	4,147	22,633	8,677	5,938
Video	211	2,105	1,687	710
App Users and Friends	**Posts**	**Tags**	**Comments**	**Likes**
Wall	3,384,740	912,687	2,152,321	1,555,225
Link	514,475	-	693,930	666,631
Photo	1,098,679	8,407,822	2,978,635	1,960,138
Video	56,241	858,054	463,401	308,763

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

- Gender
- Age
- Hometown
- Locale
- Group Memberships
- Page Likes
- Favourite Activities
- Favourite Books
- Favourite Athletes
- Favourite Teams
- Inspirational People
- Interests
- Favourite Movies
- Favourite Music
- Favourite Sports
- Favourite Television Shows
- School Information
- Work Information
- Messages data

2.3 Notation

The mathematical notation used by our classifiers during this thesis are outlined below.

define k for exposire alters

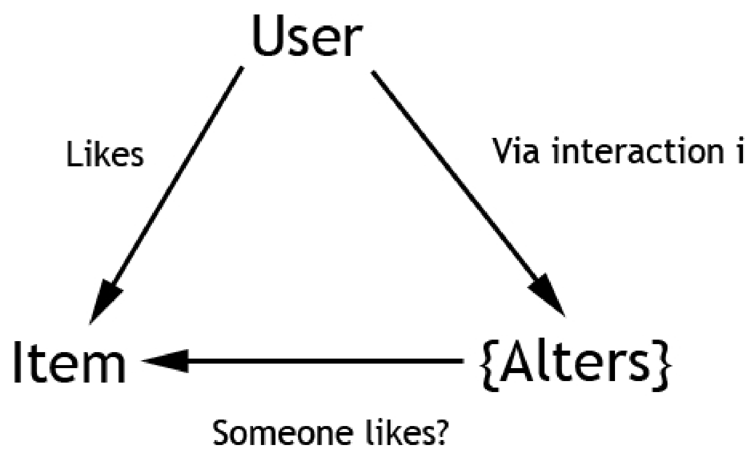

Figure 2.1: Predictors paradigm

- N users.
- M items.
- A user feature set F of size i.
- A dataset D comprised of $D = \{(n, m, f_i) \to y\}$ with the binary response $y \in \{0, 1\}$ where 0 represents a dislike and 1 represents a like.

2.4 Feature Sets

Need to discuss different feature sets

The feature sets in x can be any of the following, which are discussed further in :::

- Interactions
- Demographics
- Traits
- Groups
- Pages
- Outgoing Messages
- Incoming Messages

2.5 Previous Work

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], this work shows that user interactions are focused on a much smaller subset of their friends.

[Backstrom et al. 2011] studied two types of user uses of Facebook, explicit communication interaction and viewing attention. Communication is focused on a limited subset of friends whilst viewing attention is dispersed among a much larger set. This supports the approach of testing a wide array of user interactiuons and preferences, as each users preferences are driven by where their attention is focused.

2.5.1 Content Based Filtering

Content based filtering (CBF) [Lang 1995] is an extension of the technique of Collaborative filtering (CF) [Resnick and Varian 1997] which predicts whether a user will like an item via information about that users' prefererences as well as that of other users, CBF extends this approach by generalising from the item features which the user has explicitly liked or disliked.

Based on previous user trials [Noel 2011] Social Matchbox was the best performing algorithm in live user trials, gaining more likes then dislikes and hence will be the base line comparator used in this thesis.

2.6 Classification Algorithms

This analysis makes use of the results from a number of different classification algorithms which are outlined below.

2.6.1 Constant

The constant predictor returns a constant result irrespective of the feature vectors selected from above. The most common result in our data set is *False* and hence the *False* predictor is displayed in our analysis.

2.6.2 Social Match Box

Social Matchbox uses the Social Regularization method to incorporate the social information of the Facebook data. [Noel 2011]

What this objective component does is constrain users with a high similarity rating to have the same values in the latent feature space. This models the assumption that users who are similar socially should also have similar preferences for items.

This method is an extension of existing SCF techniques [**?**; **?**] that constrain the latent space to enforce users to have similar preferences latent representations when they interact heavily. Like Matchbox which extends regular matrix factorization methods by making use of user and link features, our extension to the Social Regularization method incorporates user features to learn similarities between users in the latent space.

2.6.3 Naive Bayes

Naive Bayes (NB) is a basic predictor which involves applying Bayes' theorem using independence assumptions between each feature in x.

The NB implementation used during this thesis is an implementation previously devised by *Scott Sanner* [**?**].

2.6.4 Logistic Regression

Logistic Regression (LR) predicts the odds of being either a like or a dislike by converting a dependent variable and one or more continuous independent variable(s) into probabability odds.

The LR implementation used during this thesis is *LingPipe* [**?**].

2.6.5 Support Vector Machine

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

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

2.7 Training and Testing

All evaluation is done using 10 fold cross validation wherein the data is partitioned into 10 complimentary subsets, each subset is composed of two separate parts one section is used for training (80%) and the other (20%) is used for testing. This is performed on 10 distinct subsets and the results are averaged across each fold.

"has been shown that positive social annotations on search items adds perceived utiltiy to the worth of a result, particularly with close social connections" [Pantel and Haas 2012]

accuracy and exposure curve

all graphs accuracy

However, if we use only a subset of the data, holding out on only elements which have at least k likes amoung the group of alters we can compare using these exposure curves which can provide some indicator of a users exposure to a liked item.

2.8 Evaluation Metrics

When evaluating the success of each method at correctly predicting the classification, the following metrics will be used.

- A *true positive* prediction refers to when the classifier correctly identifies the class as true.
- A *false positive* occurs when the prediction is true, but the true class was false.
- A *false negative* occurs when the prediction is false but the actual class is true.

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 true positive predictions divided by the sum of the true positive and false positive predictions.

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives + number of false positives}}$$

Recall refers to the number of relevant predictions that are retrieved. In the context of our results, recall refers to the number of true positive predictions divided by the sum of the true positive and false negative predicitons.

$$\text{recall} = \frac{\text{number of true positives}}{\text{number of true positives + number of false negatives}}$$

The f-score combines and balances both precision and recall and is refered to 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 in this thesis is accuracy.

Chapter 3

Interactions

The user interactions we examine in this thesis can be broken down into two distinct groups, interactions between users and messages sent between users.

3.1 User Interactions

There are a number of potential interaction mediums between users under the Facebook paradigm. These can be summarised into the following categories.

- *Direction*: The manner an interaction is received, either *Incoming* 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 *Comment*, *Tag* or *Like*.

In this case, the I for out feature vector X is defined as the crossproduct of the above components where:

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

The alters of I can then be defined as all users who have interacted with or been interacted with the current user via some i. Each component of I is set to 1 if any of the alters defined by the current set i have also liked the current item M, otherwise it is set to 0.

Applying these feature and alter sets to our classification algorthms defined above we obtain:

User interactions are marginarly outperformed by our SMB baseline, showing that user interactions do not appear to help our classification.

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.

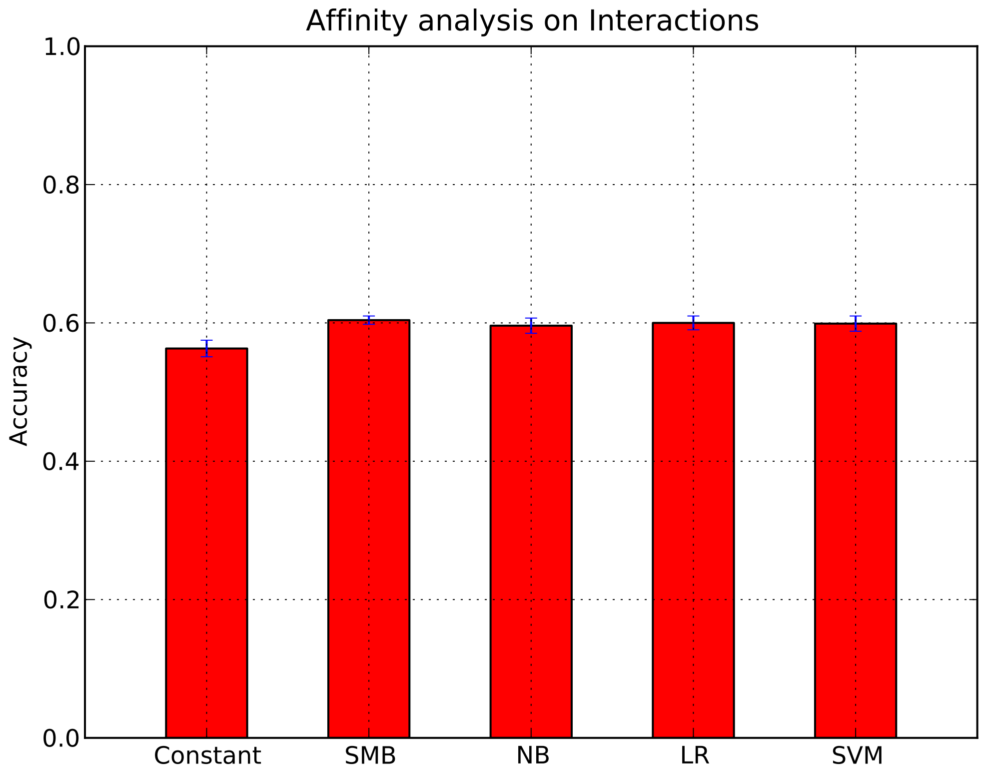

Figure 3.1: Accuracy results using the *User Interactions* feature set

Comparing *User Interactions* against our exposure curve we obtain:

We glean that as our data is restricted, the performance of our classifiers improves (note LR and SVM obtained the same results in this graph) over time. This graph shows that for *User Interactions* having one user liking an item is enough to improve upon our baselines classifiers.

3.2 Conversation

Given the nature of Facebook, it is possible for users to post or receive messages from other users.

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

Based on our LinkR data set, the most commonly used words occur with a high frequency over our user base and can be seen in the table below:

For messages the I of our feature vector X contains an element i for each of the top

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

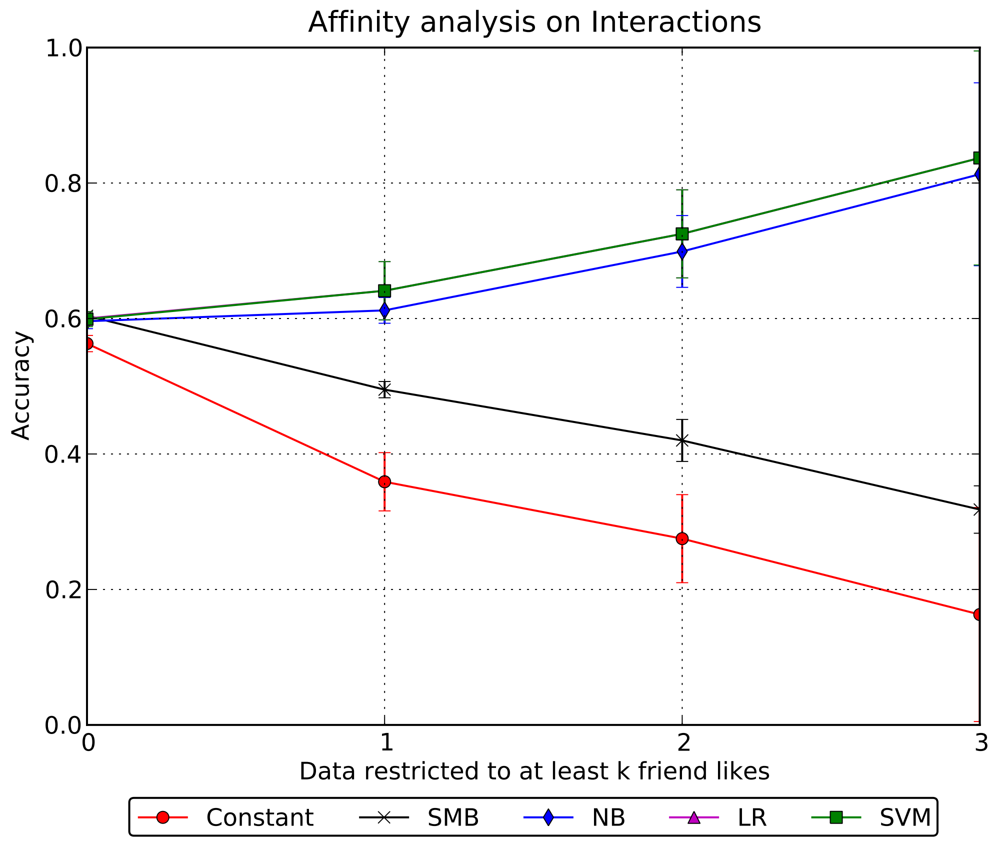

Figure 3.2: Accuracy results for an exposure curve using the *User Interactions* feature set

j most commonly used words based on the conversation content of all users.

The alters of I can then be defined as all users who have liked the current item M. Each component of I is set to 1 if any of the alters have used the current word j where $i = j$ with the user n, otherwise it is set to 0.

3.2.1 Outgoing

The first issue present is to determine the most predictive number of top words j for use by our classifiers. Given the enourmous 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: The most predictive *Outgoing Words* words sizes j for each of our classifiers are:

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

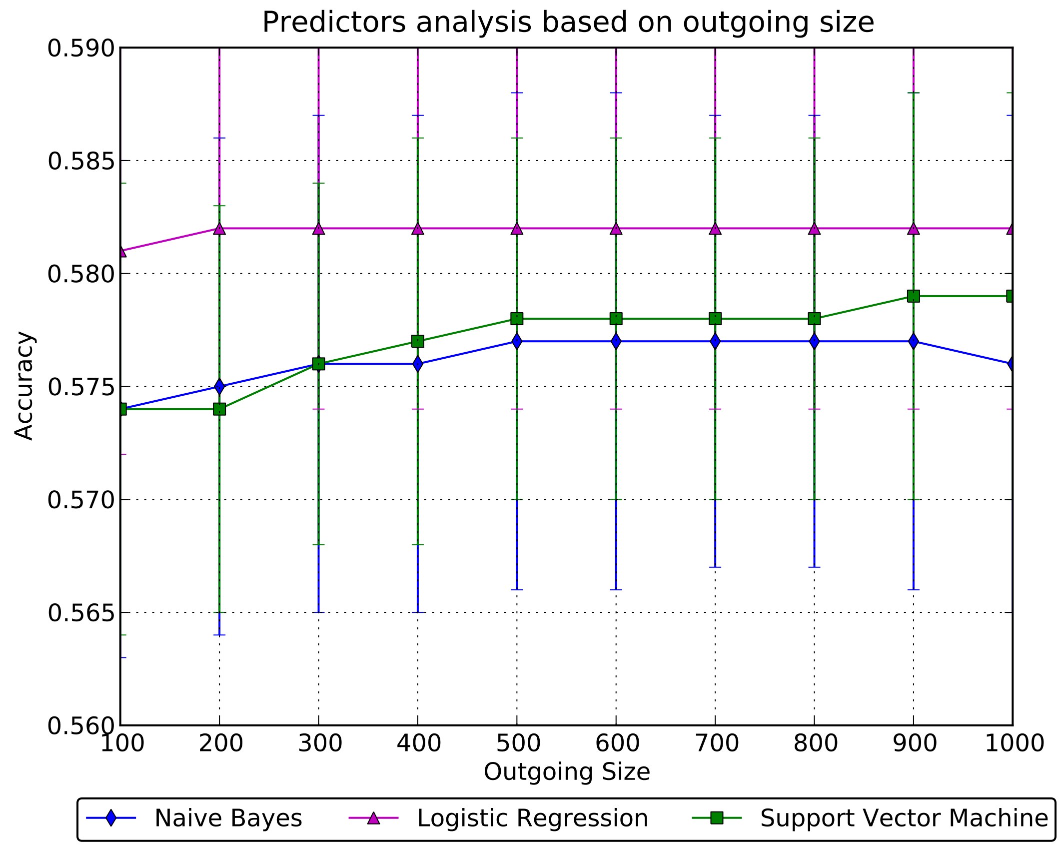

Figure 3.3: Accuracy results for different *Outgoing Words* sizes

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

These results do not show an improvement over our baselines and in fact are 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, this is elluded to in our graph above which shows an improvement in predictiveness over step sizes for SVM.

Comparing *Outgoing Messages* against our exposure curve we obtain:

Our exposure curve follows this similar trend of unimprovement for $k = 1$ likes, however as k increases there is some improvement from the baselines, but this is negligible in comparison to $k = 0$ for our classifiers.

3.2.2 Incoming

Similarly for *Incoming Words* we need to discover which is the predictive j for use of by our classifiers, using the same methodology as described above for *Outgoing Words* we obtain the following graph:

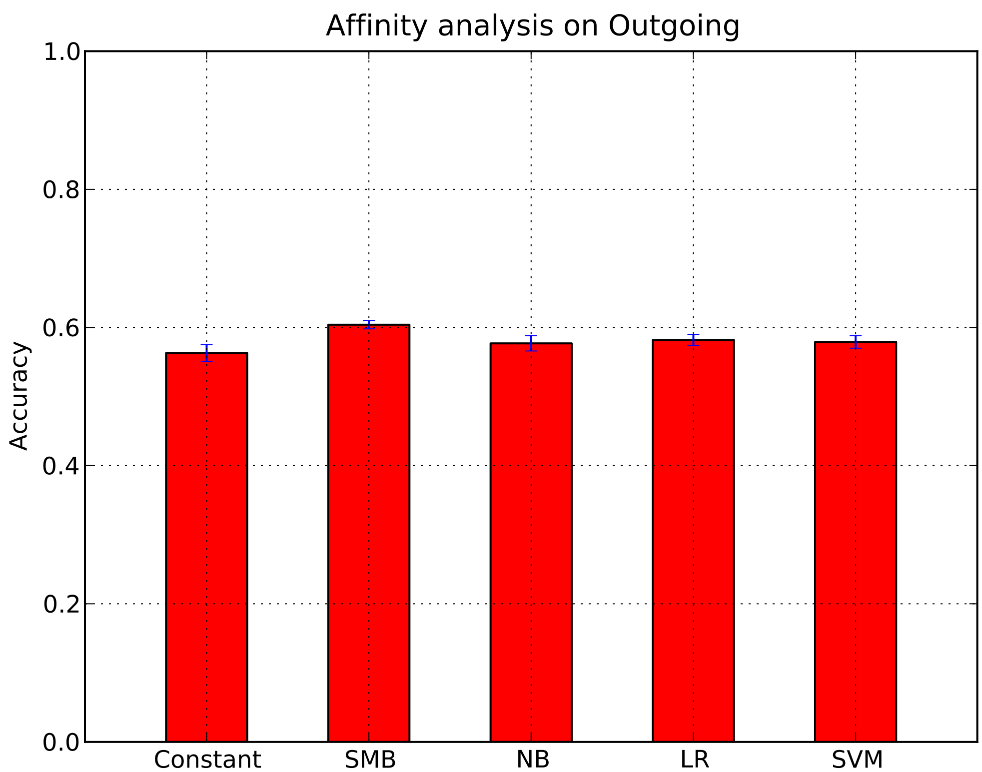

Figure 3.4: Accuracy results using the *Outgoing Words* feature set

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

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

Again, *Incoming Words* themselves are not predictive in comparison to our baselines, however not to the same extent as *Outgoing Words*.

Comparing *Incoming Messages* against our exposure curve we obtain:

Similarly, *Incoming Words* improve upon our baselines as k increases, however this performance increase is neglibible in comparison with $k = 0$ and hance *Incoming Words* do not prove to be predictive of user likes.

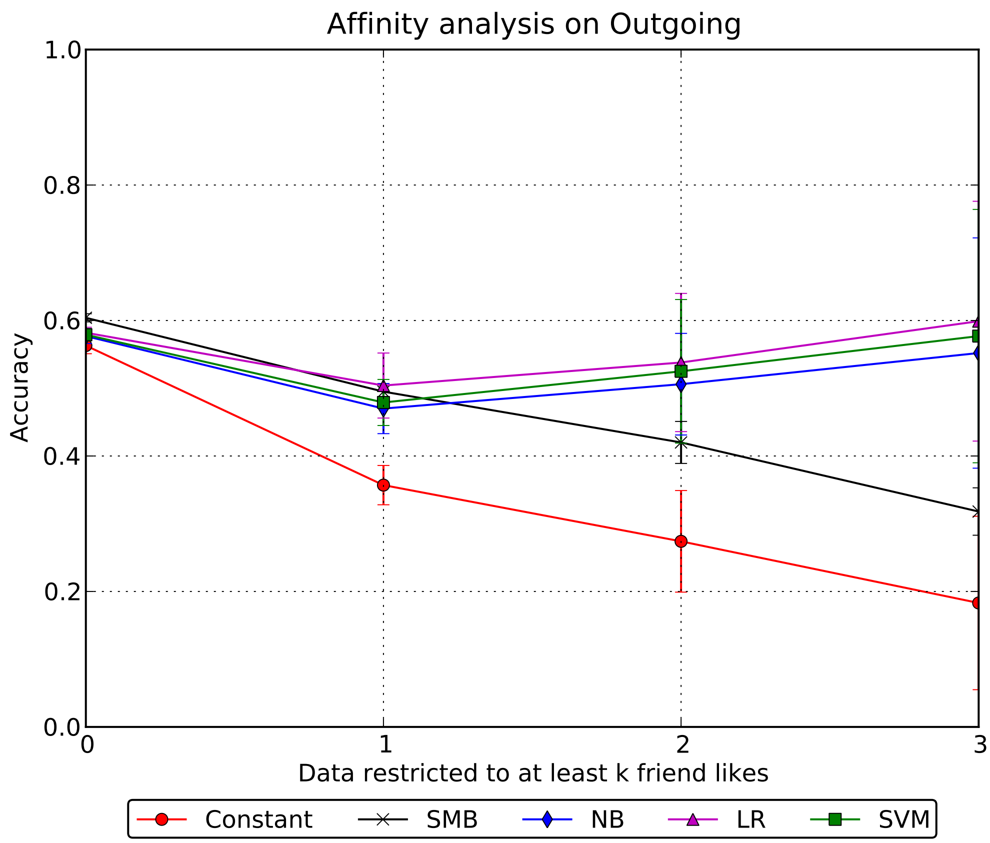

Figure 3.5: Accuracy results for an exposure curve using the *Outgoing Words* feature set

3.3 Conclusion

Throughout this section we have explored different avenues available to users to maintain interactions between other users.

We have found that words, irrespective of their directionality do not assist in improving 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 our interactions results were comparable to our baselines over $k = 0$ and this improvement continued over the exposure curve as our k increased.

[Brandtzg and Nov 2011] found that virtual interactions help reveal common interests, while real world interactions helps to suppport friendships.

Our results have shown, that for *User Interactions* it is enough for some user to have previously liked an item to allow our classification methodology to offer an increase in predictiveness as this k increases.

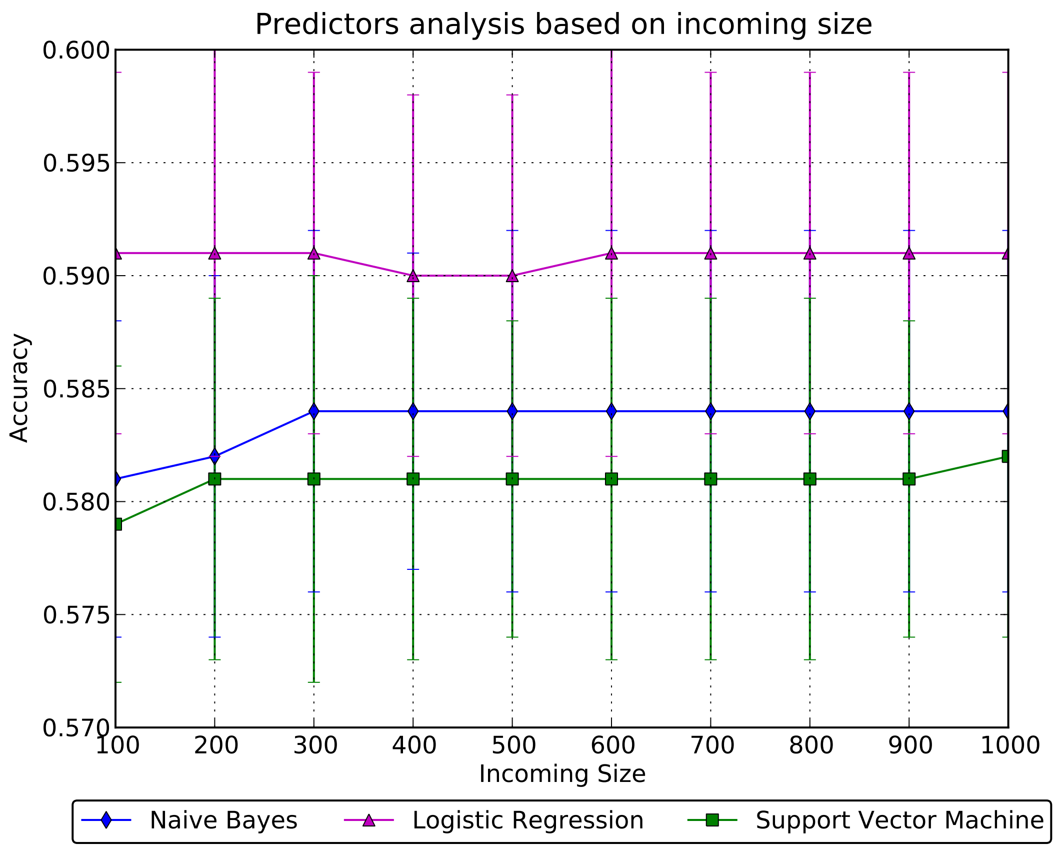

Figure 3.6: Accuracy results for different *Incoming Words* sizes

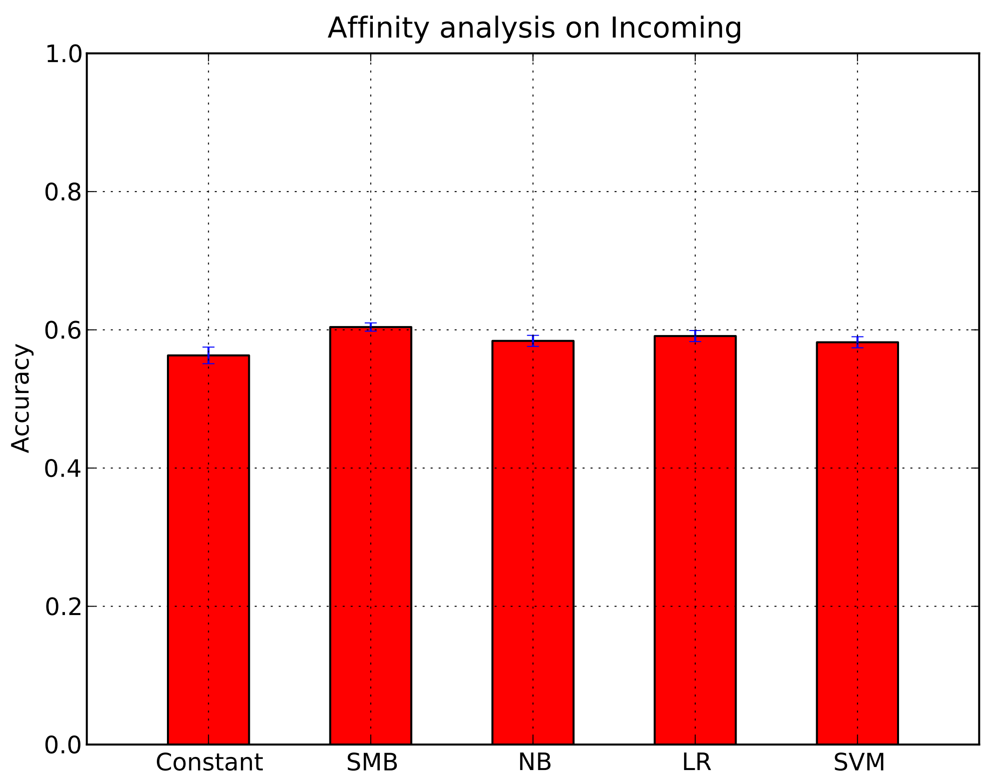

Figure 3.7: Accuracy results using the *Incoming Words* feature set

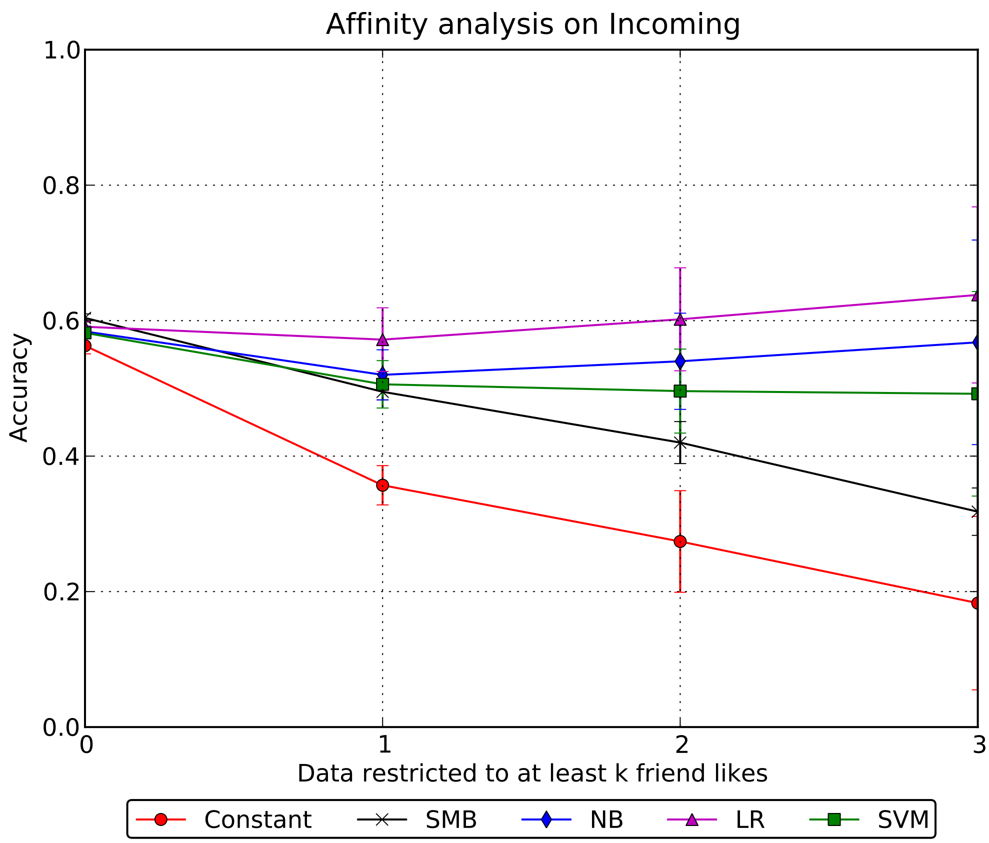

Figure 3.8: Accuracy results for an exposure curve using the *Incoming Words* feature set

Chapter 4

User Preferences

In this section we will discuss the effects of applying different types of *User Preferences* as the feature set and their predictive tendencies within our data set.

4.1 Demographics

The *Demographics* data we are interested in inlcudes:

- **Age**
- **Birthday**
- **Locale**

Below we will give a basic analysis of the *Demographics* data when extracted from our data set.

Gender breakdown:

Male	Female	Undisclosed
85	33	1

Table 4.1: Gender breakdown for app users

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 effect our results. Additionally [Backstrom et al. 2011] have shown that different genders have differing tendencies to disperse interactions across genders, implying our male skew should be unimportant. Hence gender information will be used for this feature vector.

Birthday breakdown:

Birthdays are grouped in a distinct range, most users in this data set are grouped in the age ranges of $\{18 - 30\}$. [Ugander and Marlow 2011] have found that there is a strong effect of age on friendship preferences. Hence birthday information will be used for this feature vector.

Location 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

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

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

For *Demographics* the I of our feature vector X is defined by the following conditions:

- Whether the user is male.
- Whether the user is female.
- Whether the user and any user in the alters set share the same gender.
- Whether the user and any user in the alters set share the same birthrange.

The alters of I can then be defined as the set of users who have liked the current item M. Each component of I is set to 1 if any of the alters have meet the conditions described above, in comparison (where required) with the user n, otherwise it is set to 0.

Applying this feature vector to our classifiers we obtain:

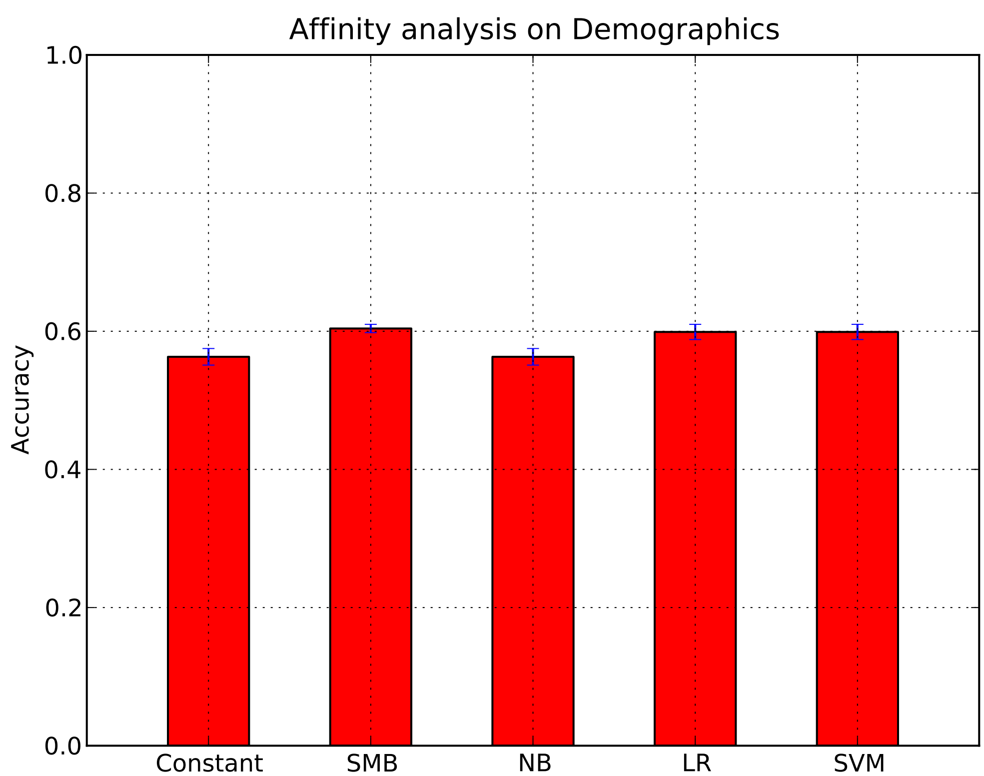

Figure 4.1: Accuracy results using the *Demographics* feature set

The *Demographics* feature vector shows our first positive results, this feature vector almost performs as well as our SMB baseline for the case of $k = 0$.

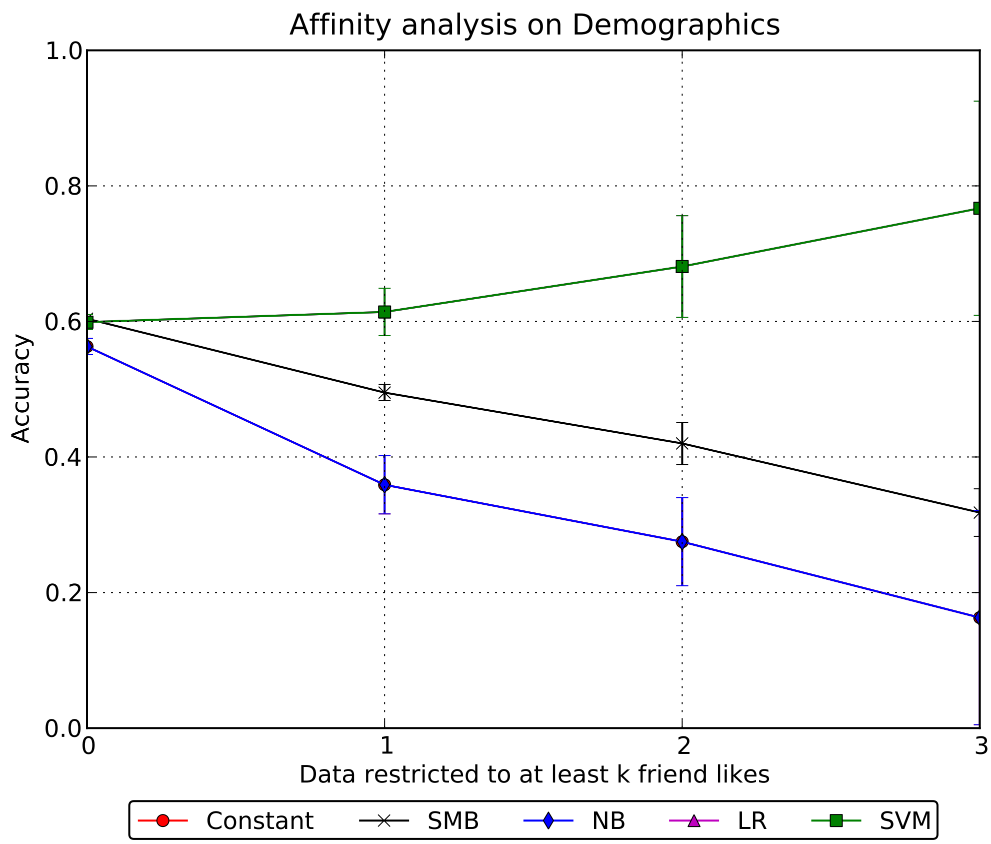

Figure 4.2: Accuracy results for an exposure curve using the *Demographics* feature set. Note in this case Constant = NB and LR = SVM

The exposure curve 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 probability that a user will like that item also increases. This positive corelation between number of likes and user likes increases with k.

4.2 Traits

Facebook facilitates a wide variety of user chosen preferences which we have defined as *Traits*. These *Traits* allow users to define under a specific area of their profile different areas or activities they are interested in.

User *Traits* data we will investigate inlcude:

- Activities
- Books

- Athletes
- Teams
- Inspirational People
- Interests
- Movies
- Music
- Sports
- Television
- School Relationships
- Work Relationships

Below we display graphs for the different *Traits* sets extracted from our data set. Followed by a subsequent analysis. Each table shows only the frequency of app users for each of the *Traits*.

The *Traits* graphed above can be broken down into three distinct sets based on their locality within the app user base.

- **High Locality**: *Music, Movies, Television* - Showing our app users appear to share similar *Traits* in a media setting.
- **Medium Locality**: *Activities, Books, Interests, Sports* - Showing our app users share some degree of similar preferences across these *Traits*.
- **Low Locality**: *Inspirational People, Athletes, Teams* - Showing our app users do not share many similar preferences across these *Traits*.

For *Traits* the I of our feature vector X is defined by the following conditions:

- Whether the current user and any user in the alters set share the same t where $T \in \{Activities, Books, Athletes, Teams, InspirationalPeople, Interests, Music, Movies, Sports, Television, Work, School\}$.

The alters of I can then be defined as the set of users who have liked the current item M. Each component of I is set to 1 if any of the alters have meet the conditions described above for each t, in comparison with the user n, otherwise it is set to 0.

Applying this feature vector to our classifiers we obtain:

The *Traits* feature vector shows our first improvement over our SMB basline in the LR and SVM case for $k = 0$ demonstrating that *Traits* are more predictive then user likes for any previously applied method.

This trend continues across the exposure curve where each successive increase of k causes the performance of our classifiers to increase.

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

This shows us that ...

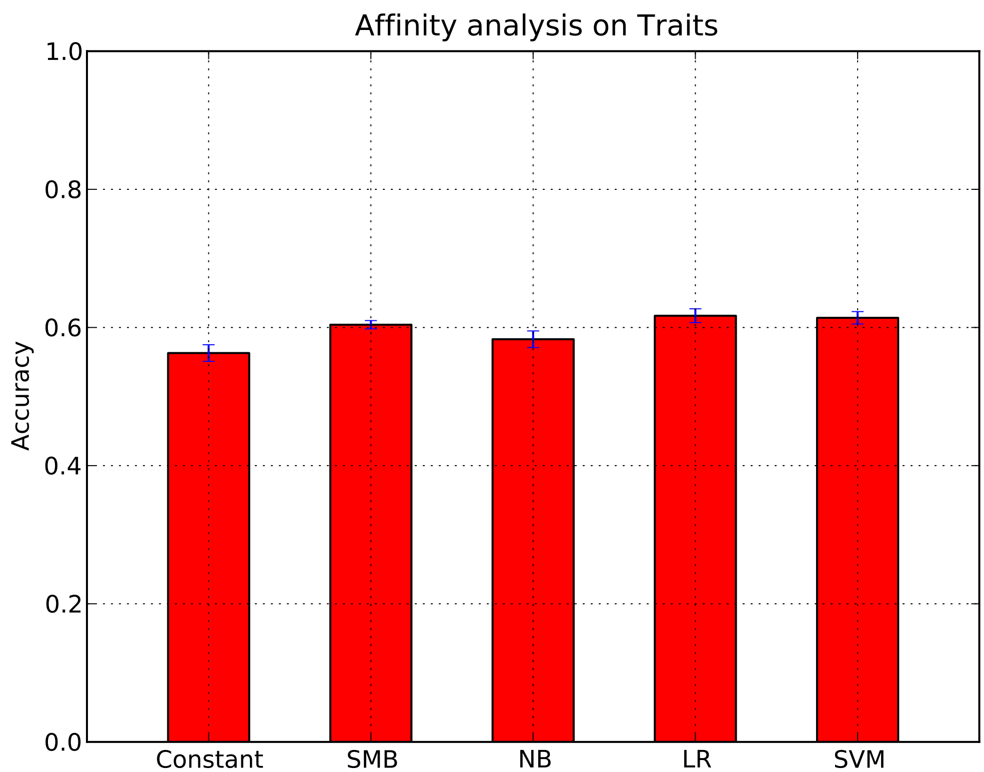

Figure 4.3: Accuracy results using the *Traits* feature set

4.3 Groups

Facebook facilitates users to join *Groups* for a large and varied set of different types ranging from local sports teams, political preferences to computer games.

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

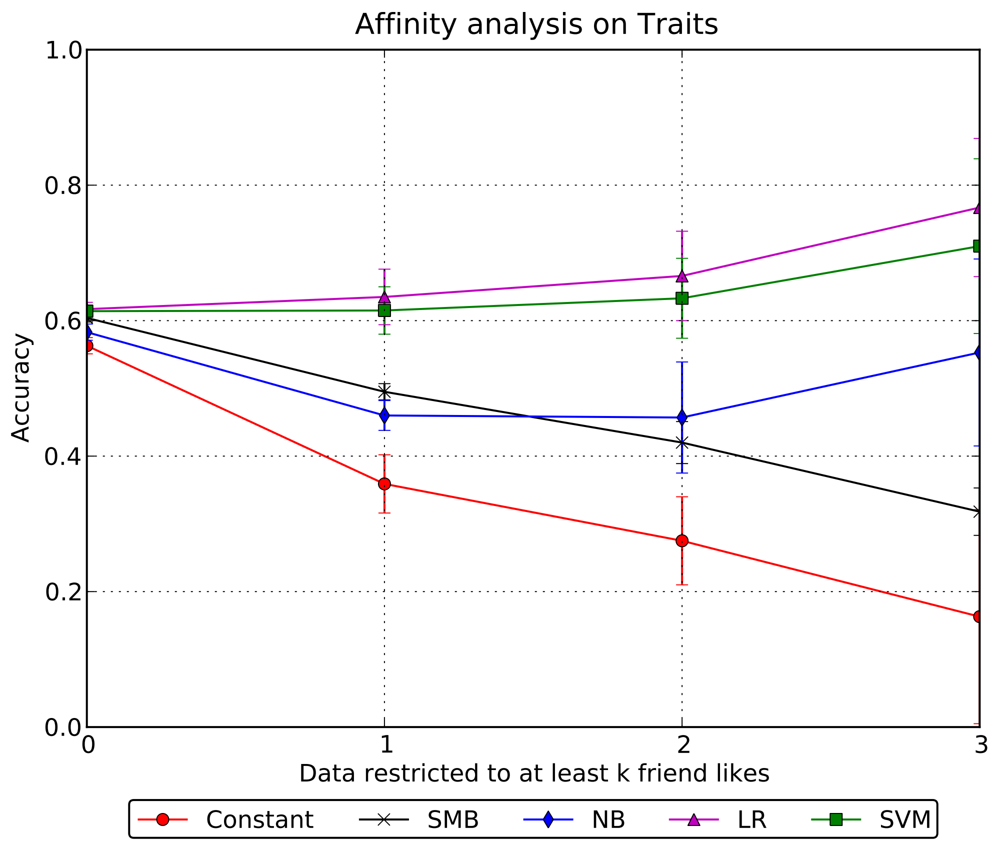

Figure 4.4: Accuracy results for an exposure curve using the *Traits* feature set

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 4.4: 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 4.5: 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 4.6: 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 entitiy.

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 4.7: 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 4.8: 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 4.9: 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 4.10: 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 4.11: 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	Brbara 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 4.12: 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 4.13: Top *Teams* for app users

Group Name	Frequency
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 4.14: App users popular *Groups* breakdown

In comparison with *Traits*, *Groups* show a higher locality for the most popular groups.

Given the quantity of groups on Facebook, we need to find some optimal 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.

The results for these tests are shown below:

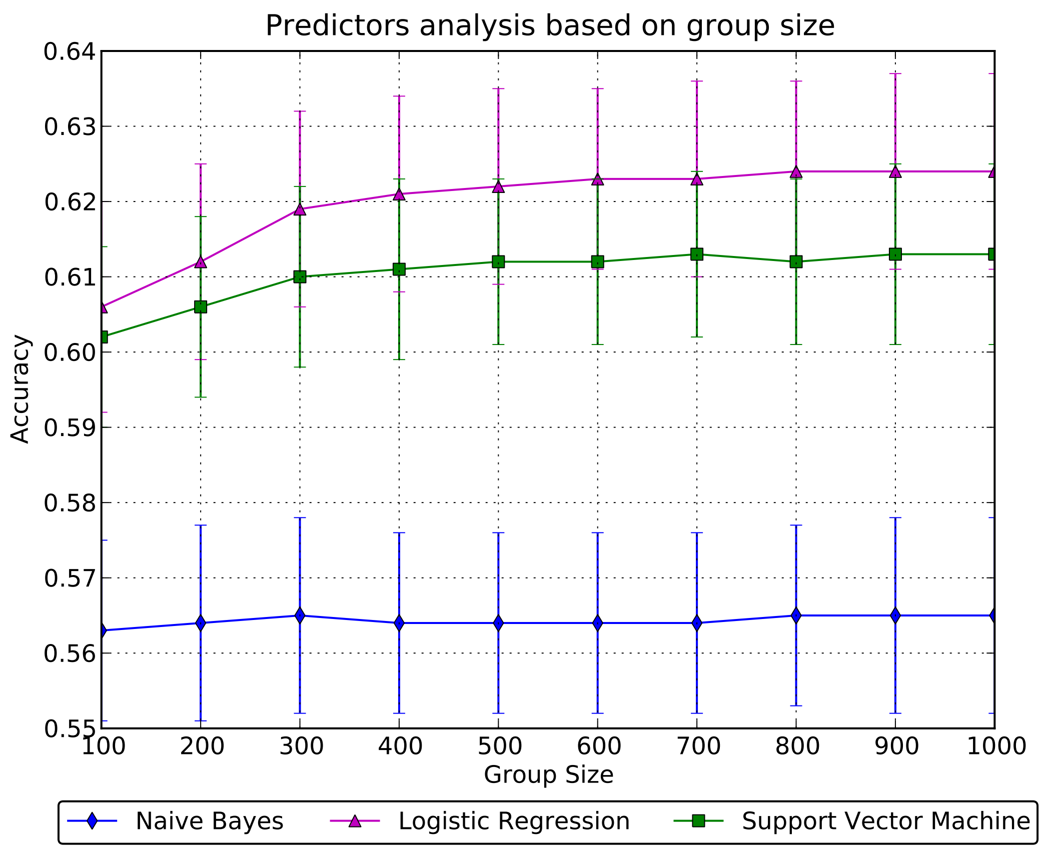

Figure 4.5: Accuracy results for different *Groups* sizes

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

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

LR and SVM show a gradual increase as this group size increases, alluding to the possibility of an even higher group size being optimal.

For *Groups* the I of our feature vector X contains an element i for each of the top j groups sizes defined above.

The alters of I can then be defined as all users who have liked the current item M. Each component of I is set to 1 if any of the alters are a member of the current group j along with the current user n, otherwise it is set to 0.

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

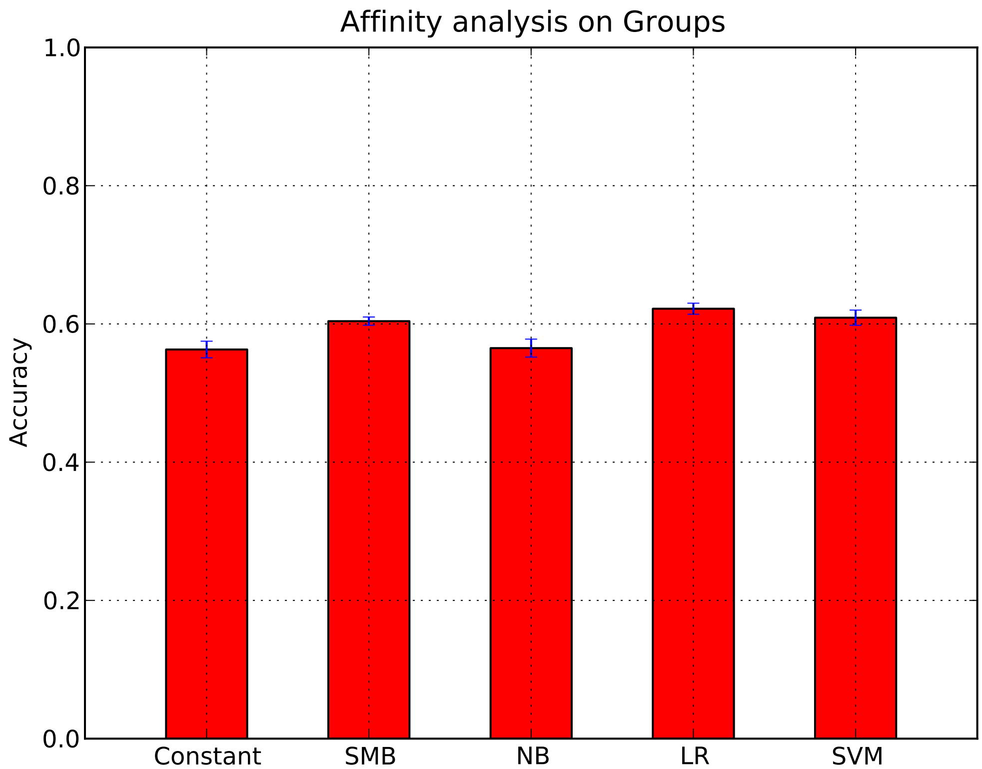

Figure 4.6: Accuracy results using the *Groups* feature set

Both LR and SVM show an improvement over our SMB baseline for the case of $k = 0$ demonstrating that *Groups* are more predictive then previous methods.

Applying the *Groups* feature vector across our exposure curve, we obtain:

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

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

This shows us that ...

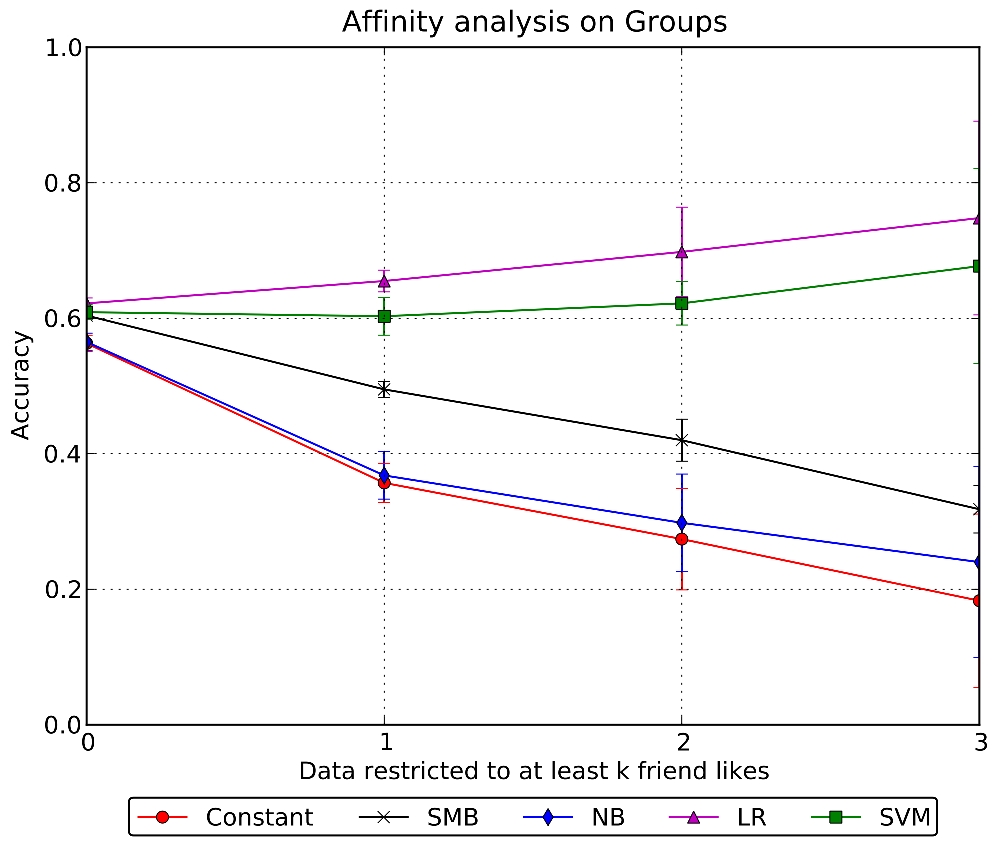

Figure 4.7: Accuracy results for an exposure curve using the *Groups* feature set

4.4 Pages

Facebook facilitates users to like *Pages* for 'things' they like across a large and varied set of different areas ranging from web browsers, TV shows to schools.

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

In comparison with *Traits* and *Groups*, *Pages* show an higher locality across the most popular pages for app users.

Given the quantity of pages on Facebook, we need to find some optimal 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.

The results for these tests are shown below:

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

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

Page Name	Frequency
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 4.15: App users *Pages* breakdown

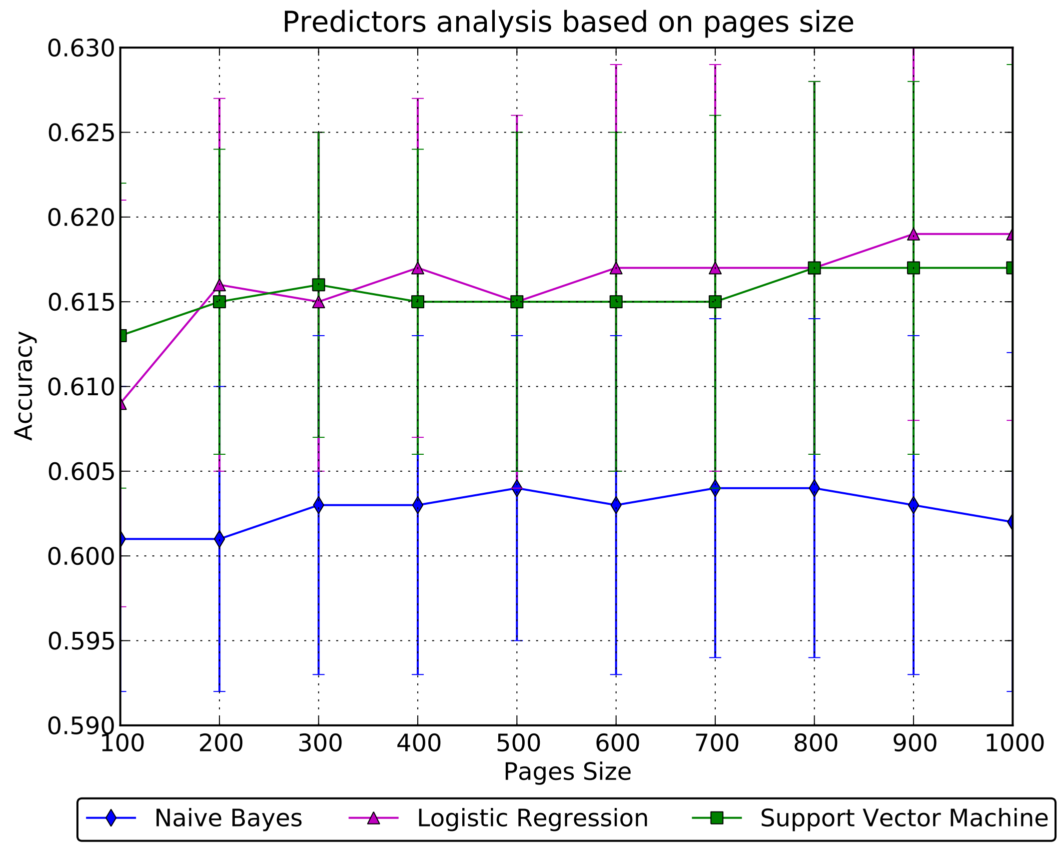

Figure 4.8: Accuracy results for different *Pages* sizes

LR and SVM show a gradual increase as this group size increases, alluding to the possibility of an even higher page size being optimal.

For *Pages* the I of our feature vector X contains an element i for each of the top j page sizes defined above.

The alters of I can then be defined as all users who have liked the current item M. Each component of I is set to 1 if any of the alters are a have liked the current page j along with the current user n, otherwise it is set to 0.

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

Both NB, LR and SVM show an improvement over our SMB baseline for the case of $k = 0$ demonstrating that *Pages* are more predictive then previous methods.

Applying the *Pages* feature vector across our exposure curve, we obtain:

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

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

This shows us that ...

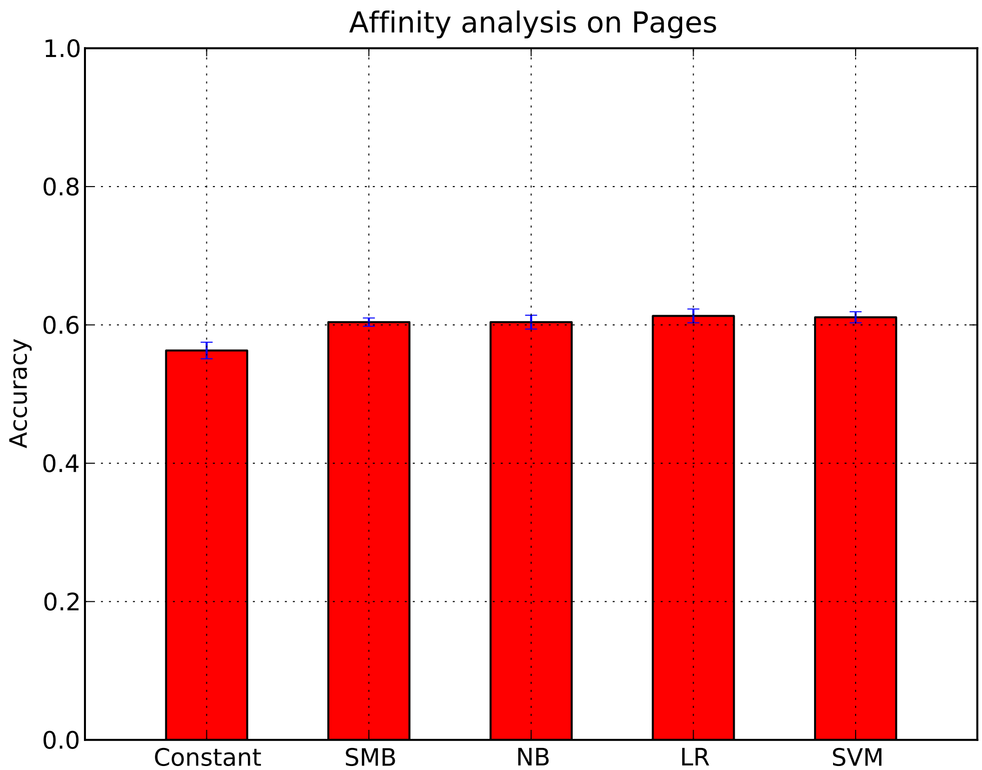

Figure 4.9: Accuracy results using the *Pages* feature set

4.5 Conclusion

Throughout this section we have explored different avenues available for users to demonstrate their personal preferences across a range of different mediums.

We have found that *User Preferences* are predictive of user likes, particularly for *Traits*, *Groups* and *Pages*. This holds true for the case of $k = 0$ and continues to improve with k.

Similarly as with *User Interactions*, our results have shown, that it is enough for some user to have liked an item to allow our classification methodology to offer an increase in predictiveness.

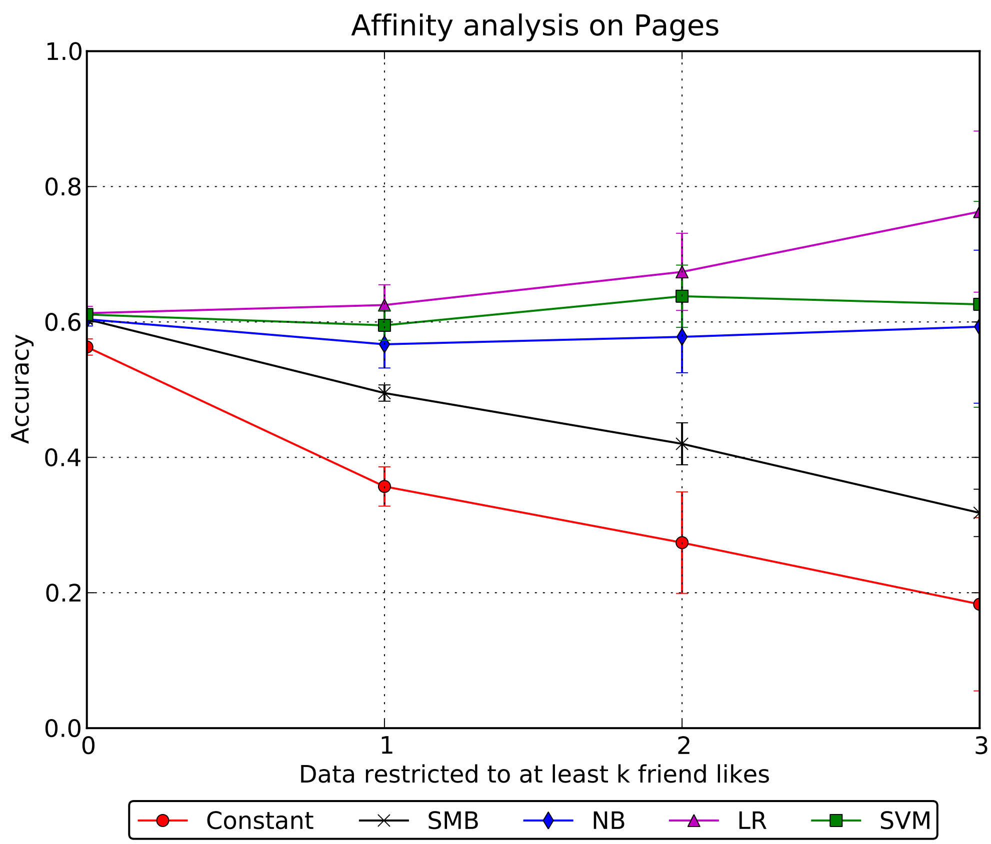

Figure 4.10: Accuracy results for an exposure curve using the *Pages* feature set

Chapter 5

Feature Combinations

As outlined above, features which positively contributed to our classifiers were from the *Traits*, *Groups* and *Pages* feature vectors. In this section we combine these positive feature vectors together into a larger feature vector comprised of the individual positively contributing elements.

5.1 Positive Feature Selection

Using the combined feature vector X where I is comprised of:

- Traits
- Groups
- Pages

Applying this feature vector to the data set:

We find that the *Combination* feature vector gives better results for our classifiers when compared with our baselines.

Classifier	**Accuracy**
SMB	0.604 ± 0.006
NB	0.583 ± 0.012
LR	0.617 ± 0.01
SVM	0.614 ± 0.009

Table 5.1: *Traits* results for $k = 0$

Classifier	**Accuracy**
SMB	0.604 ± 0.006
NB	0.604 ± 0.01
LR	0.613 ± 0.01
SVM	0.611 ± 0.008

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

Classifier	**Accuracy**
SMB	0.604 ± 0.006
NB	0.565 ± 0.013
LR	0.622 ± 0.008
SVM	0.609 ± 0.011

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

Classifier	**Accuracy**
SMB	0.604 ± 0.006
NB	0.605 ± 0.01
LR	0.624 ± 0.009
SVM	0.618 ± 0.01

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

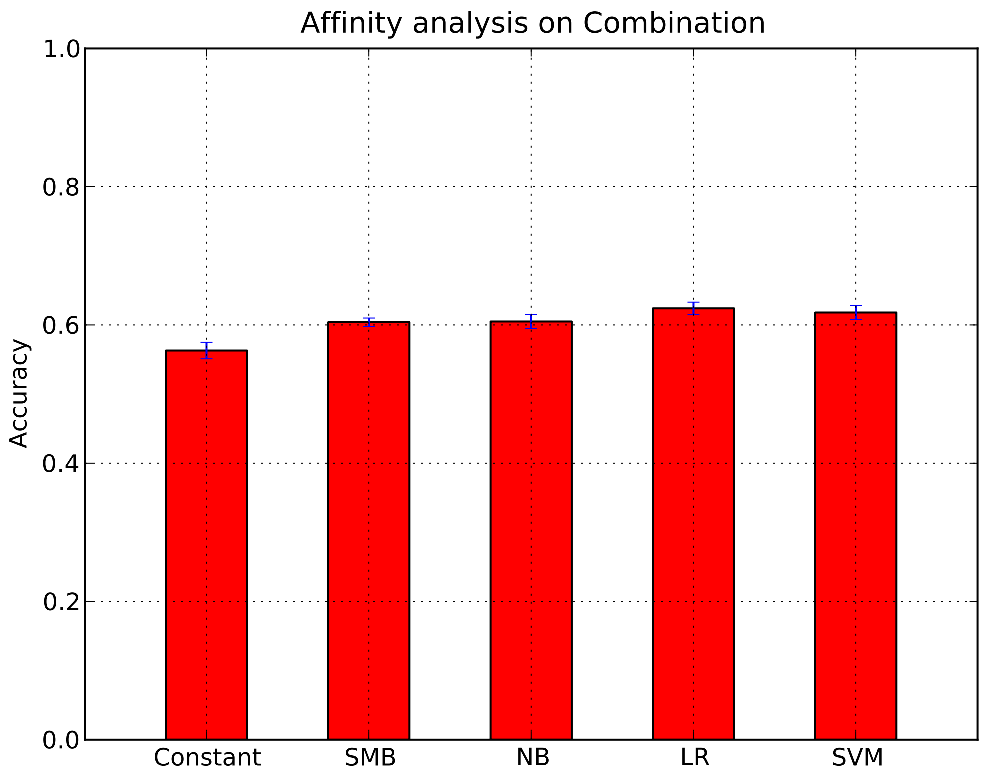

Figure 5.1: Accuracy results using the *combined* feature set

This shows that based on our results the most predictive feature vector is a combination of the individual best feature vectors found in our *User Preferences* section.

This trend continues over the exposure curve with LR, SVM and NB all improving as k increases.

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

This shows us that ...

5.2 Summary

In this thesis we have tested and compared different feature vectors across different exposures of size k. We have shown that *User Interactions* in themselves are not predictive of user likes, however coupled with the likes exposure curve, they do show an improvement over our baselines.

We have also shown the interesting result that *User Preferences* are predictive of user likes in the base case of $k = 0$ and this trend continues over the likes exposure

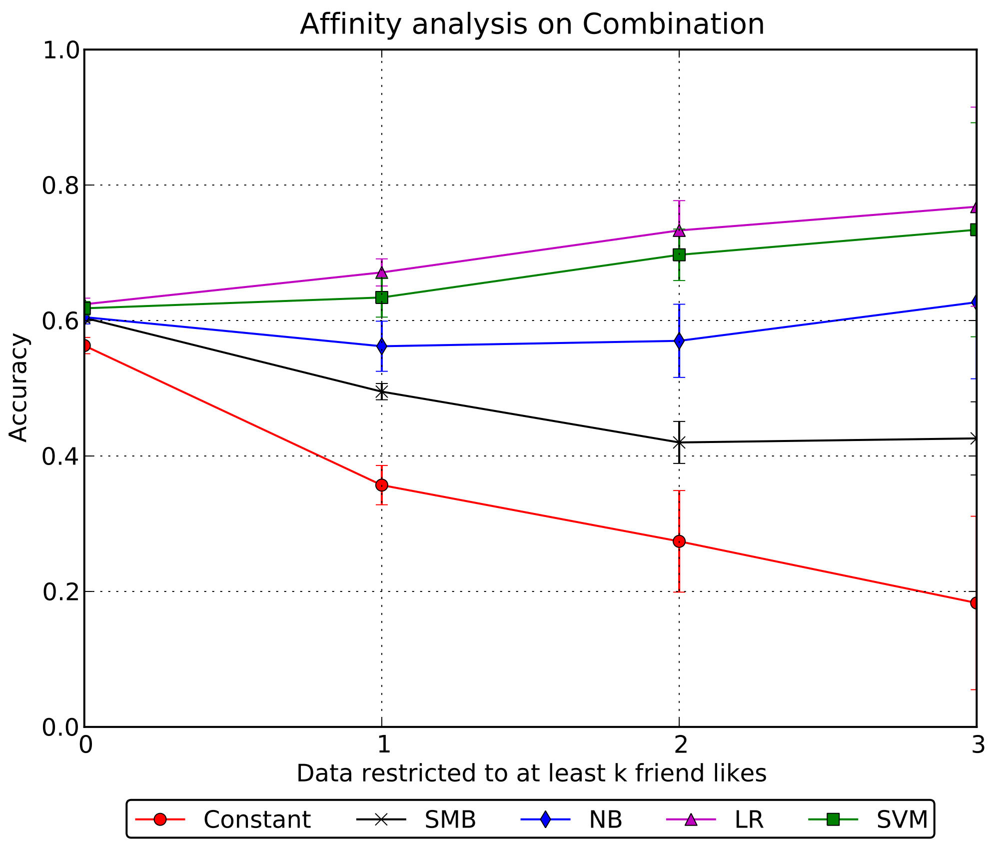

Figure 5.2: Accuracy results for an exposure curve using the *pages* feature set

curve as k increases.

To answer the question initially proposed for this thesis, we have shown the feature vector which provides the highest predictiveness for user likes is the combination of the individually predictive feature vectors which is the novel insight provided by this thesis.

5.3 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 size for each classifier

- **Ranges tests**: *Groups* and *Pages* were tested based on their popularity, a more useful comparison could be to test based on some other condition, such as *Group* or *Page* size.

- **Individual Traits analysis**: During our traits analysis the feature vector was set to 1 if the user and any user in the set of alters were part of the same *Traits* group, it would be beneficial to do an individual analysis on each component of the *Traits* data to find which parts of *Traits* are most predictive (similar analysis as done in *Groups*).

- **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 hence why the LinkR 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.

- **Cold start**: Leaving out some subset of users when training our models, but including them during testing.

- **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 indicitive of the general Facebook user population would result in more generalsied results.

- **Bayesian Model Averging**: Weighting the most successful machine learning models under different feature sets and exposure curves to simulate a new combined classifier.

Bibliography

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. 15)

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). (pp. 6, 19)

BOND, J. K. M. S., FARISS AND FOWLER. 2011. A 61-million-person experiment in social influence and political mobilization. *Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. 489*.

BRANDTZG, P. B. AND NOV, O. 2011. Facebook use and social capital — a longitudinal study. ICWSM'11 (2011). (p. 15)

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. 7)

GRANOVETTER, M. S. 1978. Threshold models of collective behavior. *Am. J. Sociol 83(6):14201443*. (p. 1)

HILL, R. AND DUNBAR, R. 2003. Social network size in humans. *Human Nature 14*, 1, 53–72. (p. 6)

LANG, K. 1995. NewsWeeder: Learning to filter netnews. In *12th International Conference on Machine Learning ICML-95* (1995), pp. 331–339. (p. 6)

NOEL, J. G. 2011. New social collaborative filtering algorithms for recommendation on facebook (2011). (p. 6)

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. 7)

RESNICK, P. AND VARIAN, H. R. 1997. Recommender systems. *Communications of the ACM 40*, 56–58. (p. 6)

ROMERO, D. M., MEEDER, B., AND KLEINBERG, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. WWW '11 (2011), pp. 695–704. ACM. (p. 1)

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. 9)

UGANDER, B., KARRER AND MARLOW. 2011. The anatomy of the facebook social graph. *CoRR abs/1111.4503*. (pp. 19, 42)

WATTS, D. J. AND STROGATZ, S. H. 1998. Collective dynamics of small-world networks. *Nature 393(6684):440442*. (p. 1)