# 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 2ε.

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

Riley Kidd
11 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 also compared to previous work using *Matchboxing and Social Matchboxing* techniques. 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

# 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. Providing a myriad of expanding social information and user driven content.

Our question becomes, how can we exploit this information to decipher which user interactions or preferences are most indicitive of user likes? Can we leverage this information and generalise over a social population?

Internet is becoming a network of people

## 1.1 Objectives

The primary objective of this thesis is to contrast and compare differing user feature sets. Using the machine learning concepts of *Naive Bayes*, *Logistic Regression* and *Support Vector Machines* compared with our appropriate baslines of *Social Mathbox* and *Constant Classifiers*.

Using these classification techniques we will compare both user interaction features *interactions, incoming/outgoing messages* and user preference features *demographics, traits, groups, pages*. To analyse the effect each feature set has on classification.

These algorithms will also be tested against a subset of the data, using a hold out technique to only test against items which has been liked by some friend.

Finally, we will analyse the effect of combining successful user feature sets together and the results of combining successful models together using a *Bayesian Model Averaging* approach.

## 1.2 Contributions

In the preceding section, we outlined different user interaction and preference features. Our specific contributions made during this thesis show

"Facebook users during the 2010 US congressional elections. The results show that the messages directly influenced political self-expression, information seeking and

real-world voting behaviour of millions of people. Furthermore, the messages not only influenced the users who received them but also the users friends, and friends of friends. The effect of social transmission on real-world voting was greater than the direct effect of the messages themselves, and nearly all the transmission occurred between close friends who were more likely to have a face-to-face relationship. These results suggest that strong ties are instrumental for spreading both online and real-world behaviour in human social networks." [**?**]

discuss feed time for facebook

"Social inuence can play a crucial role in a range of behav- ioral phenomena, from the dissemination of information, to the adoption of political opinions and technologies [23, 42]," [**?**] [**?**]

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

- Both *interactions* and *incoming/outgoing messages* are not more predictive then previously used *Social Matchboxing* techniques.
- Each user preference of *demographics*, *traits*, *groups*, *pages* contributed to a better result then our baselines.
- Combining user preferences with a hold out technique for user likes results in a sgnificant improvement from our baslines.
- Combination of user preferences is better
- Model combination can encapsulate these ideas

Overall, we provide a methodology which improves upon previous work and offers a way to combine positively contributing aspects of different feature sets 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 the data set, 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 sets for user interactions and the results from applying these feature sets in comparison with our baselines.
- **Chapter 4**: Similarly as above, however we discuss user preferences
- **Chapter 5**: In this chapter we discuss results from combining different feature sets and models

- **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 which feature sets are predictive of user likes and offer an approach to combining positive components into a useful classifier.

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 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.

## 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:

- Gender

[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.

- 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.

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

## 2.4 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) [?] is an extension of the technique of Collaborative filtering (CF) [?] 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.

Baesd on previous user trials [?] Social Matchbox was the best performing algorithm in live user trials, gaining more likes then dislikes.

## 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

### 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.

## 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

One reason could be we can not track information passing outside of Facebook, users who frequently interact could be real world friends who share information bia IM or email.

There are a number of potential interactions between users under the Facebook paradigm.

- Direction: Incoming or outgoing Given the approach of [Saez-Trumper et al. 2011] where interaction directionality was shown to be highly reflective of user preferences we decipher between incoming and outgoing interactions
- Modality: Links, posts, photos or videos
- Type: Comment, tag or like

In this case, our feature vector $x_i$ is comprised of the crossproduct, where:

$$i = \{incoming, outgoing\} \times \{post, photo, video, link\} \times \{comment, tag, like\}$$

The alters of $i$ can then be defined as all users who have interacted with the current user via some interaction $i$. The column is set to 1 if any of the alters defined by the current set $i$ have also liked the item associated with the user, otherwise it is set to 0.

Applying our classification algorthms defined above and this feature set we obtain:

Compared with our baselines, user interactions do not appear to help our classification.

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 [Romero et al. 2011] which can provide some indicator of a users exposure to a liked item.

Having one user liking an item is simpley enough.

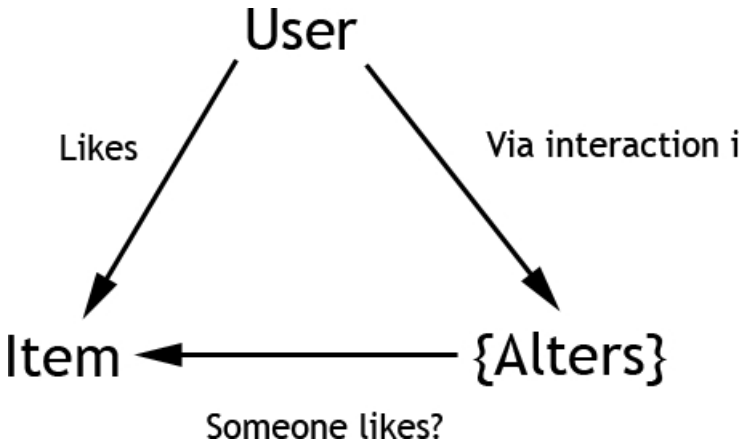


**Figure 3.1**: Predictors paradigm

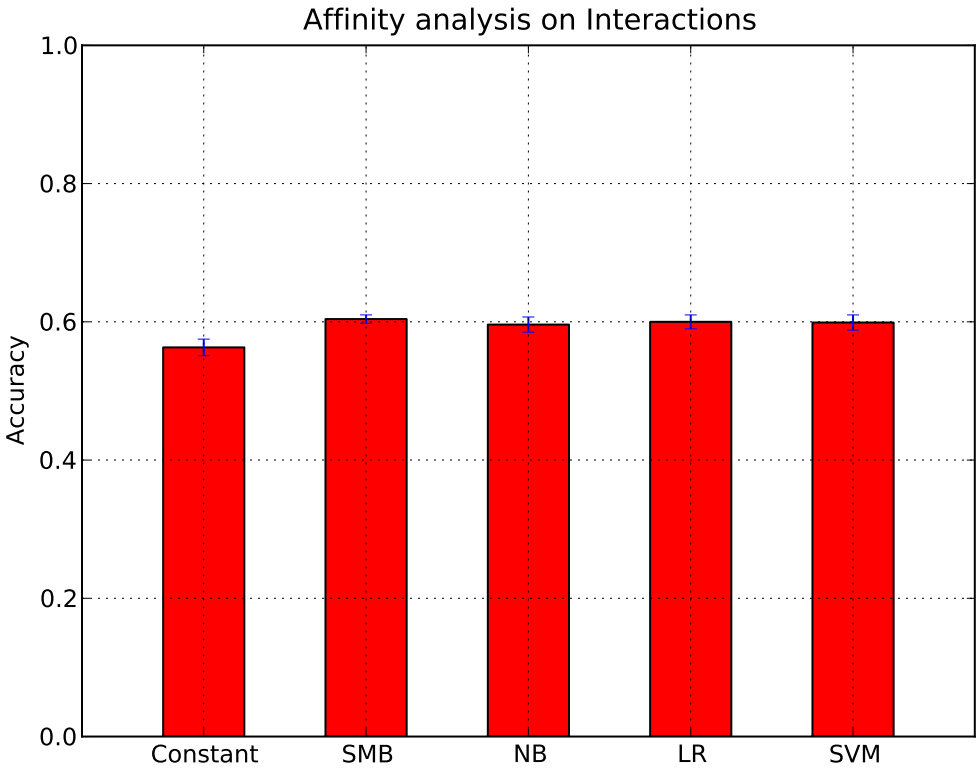


**Figure 3.2**: Accuracy results using the *user interactions* feature set

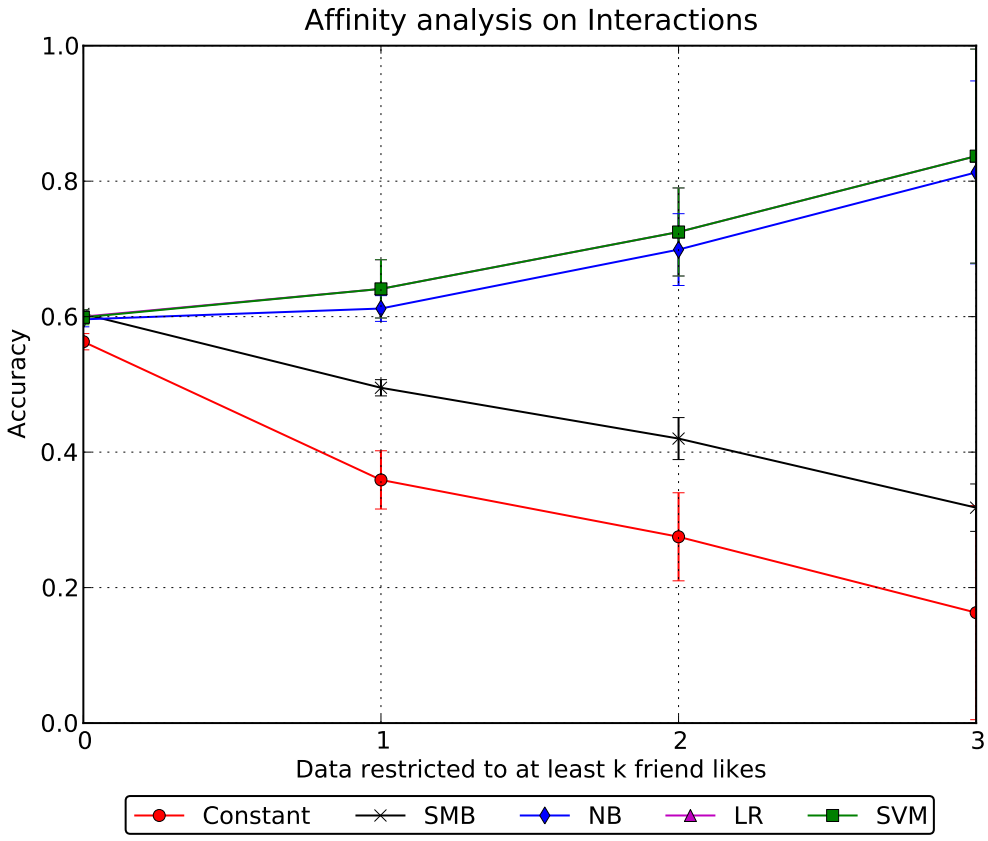


**Figure 3.3**: Accuracy results for an exposure curve using the *user interactions* feature set

## 3.2 Conversation

Given the nature of Facebook, it is possible for users to post messages to 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.

The most commonly used words are 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 |
| 41 | right | 52,112 |
| 42 | best | 51,941 |
| 43 | awesome | 51,663 |
| 44 | hope | 50,980 |
| 45 | 2 | 50,720 |
| 46 | next | 50,375 |
| 47 | work | 49,459 |
| 48 | way | 49,358 |
| 49 | man | 49,101 |
| 50 | :( | 48,184 |
| 51 | ¡3 | 47,985 |
| 52 | even | 47,480 |
| 53 | 4 | 46,068 |
| 54 | us | 45,919 |
| 55 | pretty | 44,804 |
| 56 | hey | 44,614 |
| 57 | say | 44,315 |
| 58 | better | 43,357 |
| 59 | thanx | 42,639 |
| 60 | bro | 41,187 |
| 61 | take | 41,081 |
| 62 | always | 40,457 |
| 63 | wow | 40,452 |
| 64 | pic | 40,185 |
| 65 | though | 40,032 |
| 66 | actually | 39,565 |
| 67 | last | 39,175 |
| 68 | thats | 38,833 |
| 69 | cool | 37,844 |
| 70 | dear | 37,328 |
| 71 | ok | 36,441 |
| 72 | sorry | 36,345 |
| 73 | never | 36,000 |
| 74 | thing | 35,941 |
| 75 | first | 35,785 |
| 76 | looks | 35,496 |
| 77 | night | 35,475 |
| 78 | thought | 34,458 |
| 79 | photo | 33,989 |
| 80 | & | 33,902 |

**Table 3.1**: Top conversation content data for all users

### 3.2.1 Outgoing

First we need to figure out the number of outgoing messages words are optimal for our classifiers. Given the size of potential messages we decided to test within a range of (100-1000) where each word is based on its top frequency.

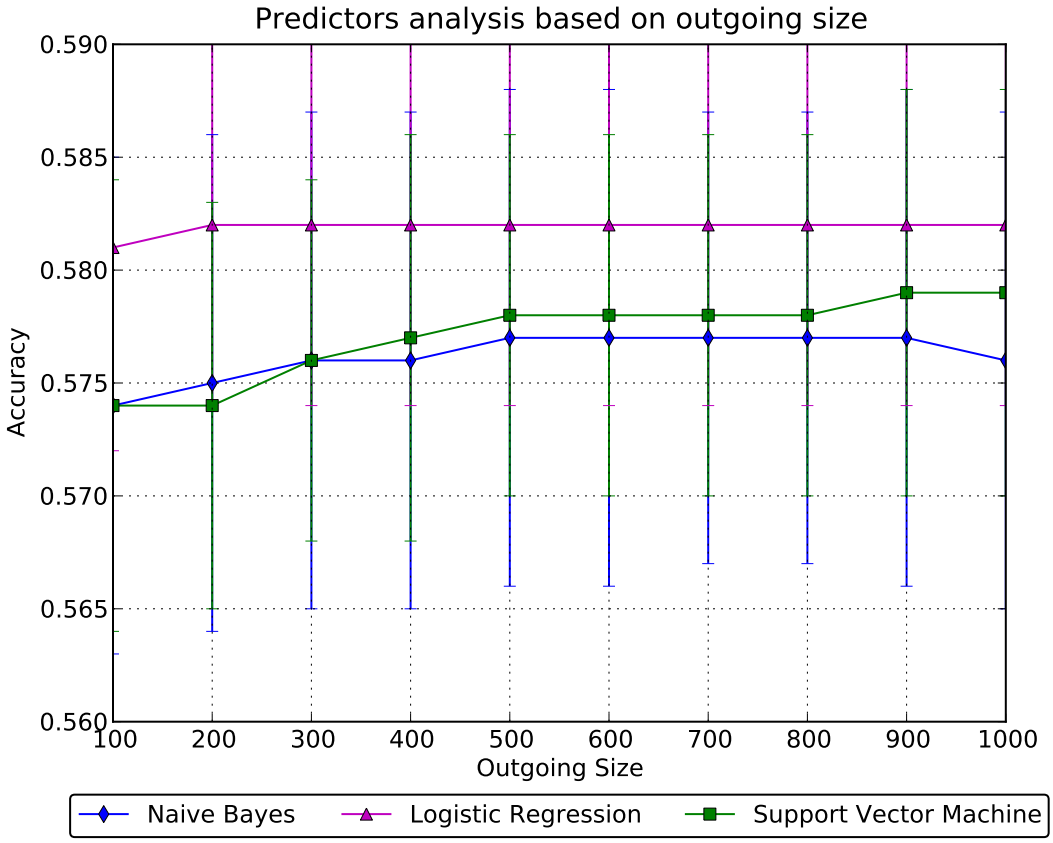


**Figure 3.4**: Accuracy results for different *outgoing words* sizes

The most predictive top outgoing words sizes for our classifiers are:

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

Testing our classifiers using their optimal $x_i$ based on the above analysis we obtain:
These results do not show an improvement over our baselines.

However when using an exposure curve we find that having some user liking our item is enough to improve classification.

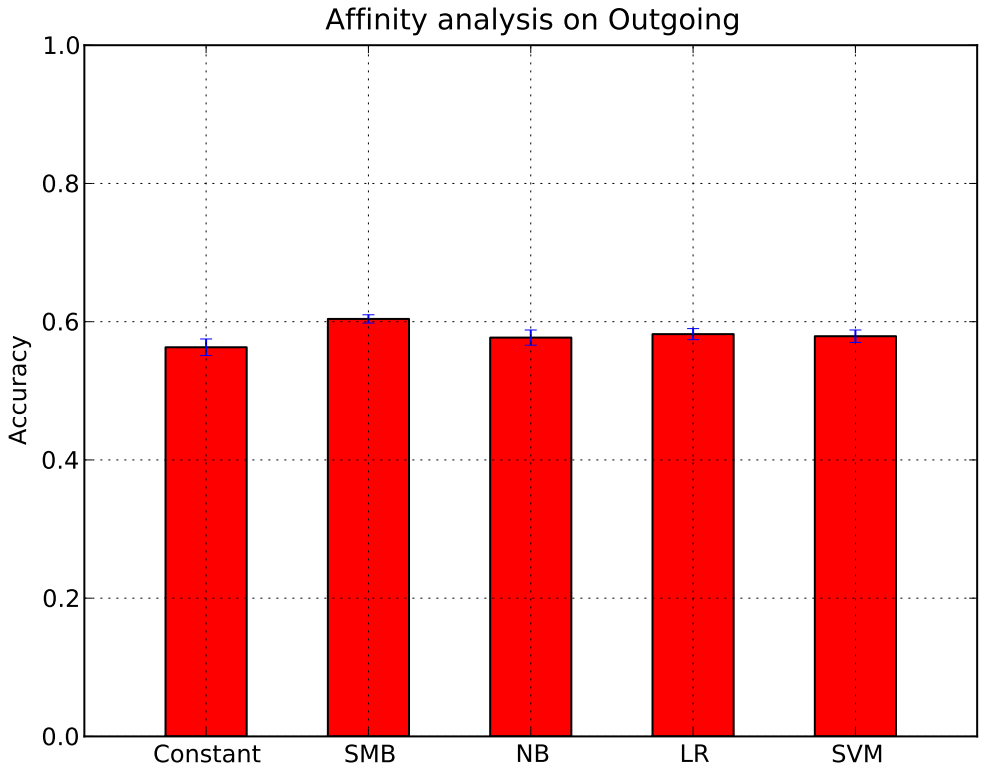


**Figure 3.5**: Accuracy results using the *incoming words* feature set

### 3.2.2 Incoming

Similarly for incoming messages we need to discover which is the best.

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

Testing using these optimal message size groups we obtain.
Again incoming messages themselves are not predictive.

## 3.3 Conclusion

User interactions in themselves are not predictive of user likes.

[**?**] concluded that it is less important what users say, then who they interact with.

This is supported by [**?**] who found that virtual interactions help reveal common interests, while real world interactions helps to suppport friendships.

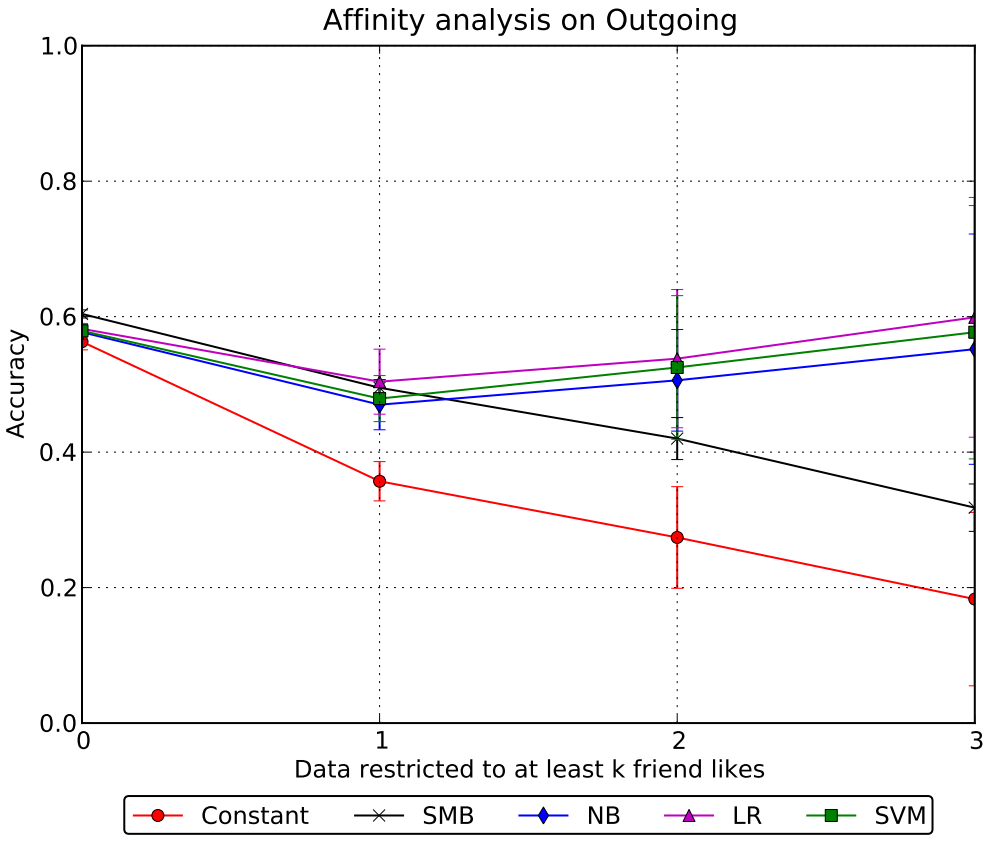


**Figure 3.6**: Accuracy results for an exposure curve using the *outgoing words* feature set

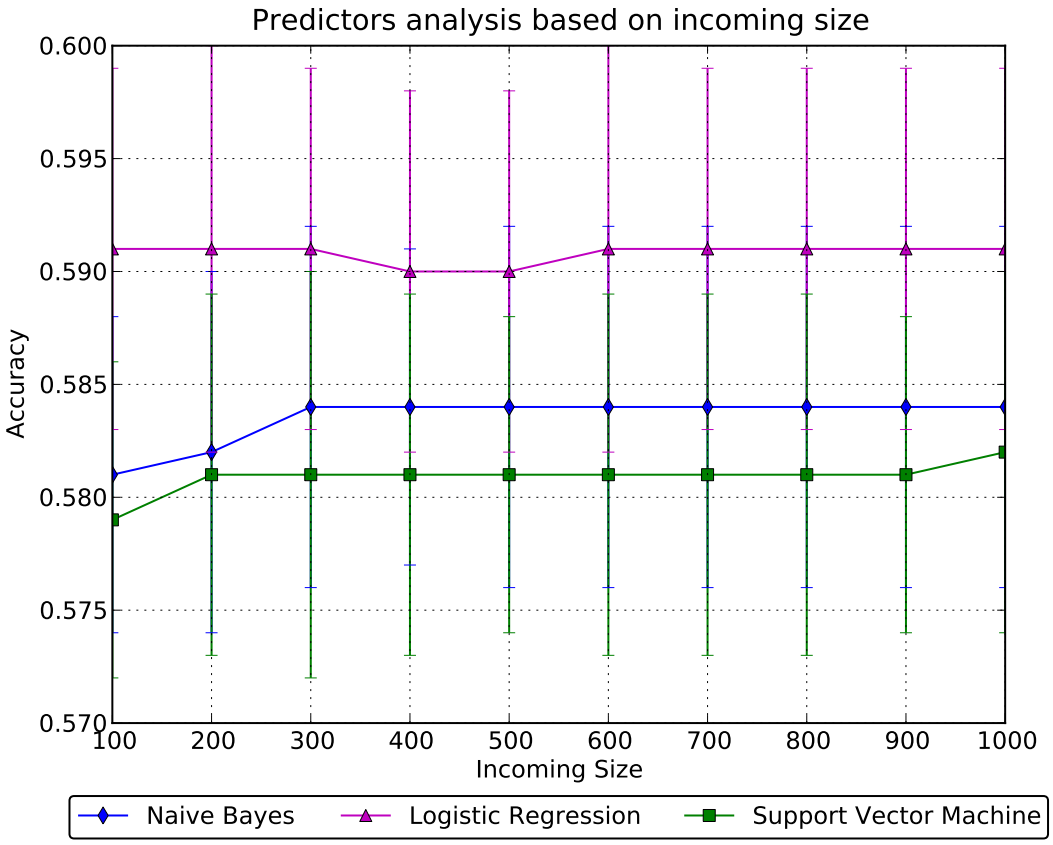


**Figure 3.7**: Accuracy results for different *incoming words* sizes

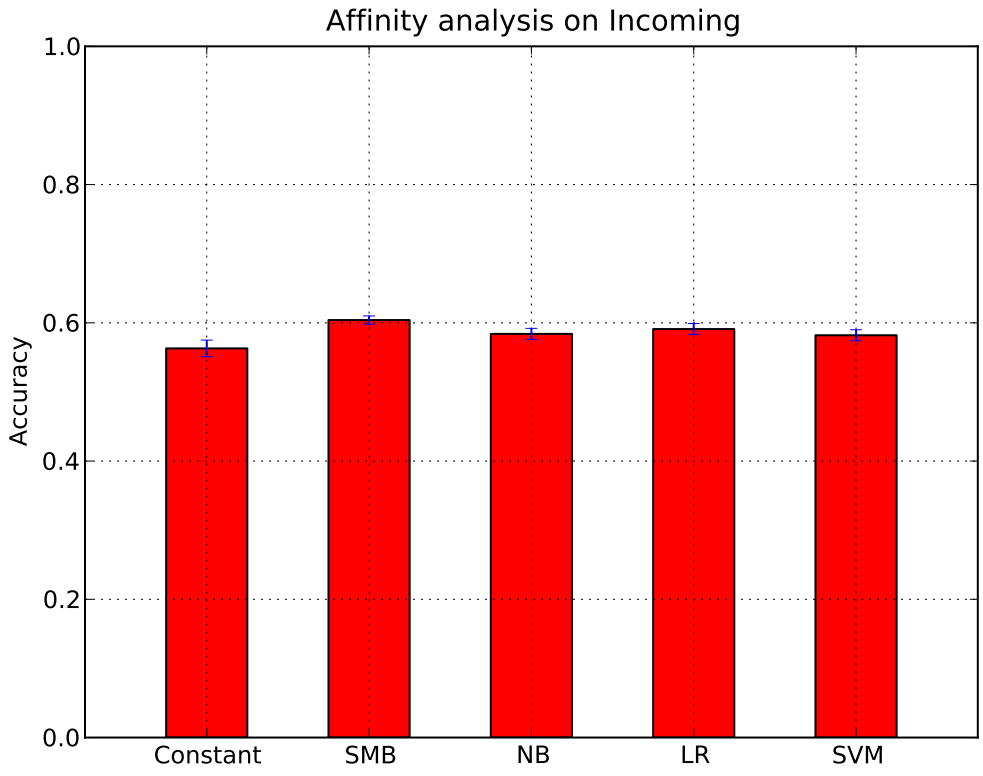


**Figure 3.8**: Accuracy results using the *incoming words* feature set

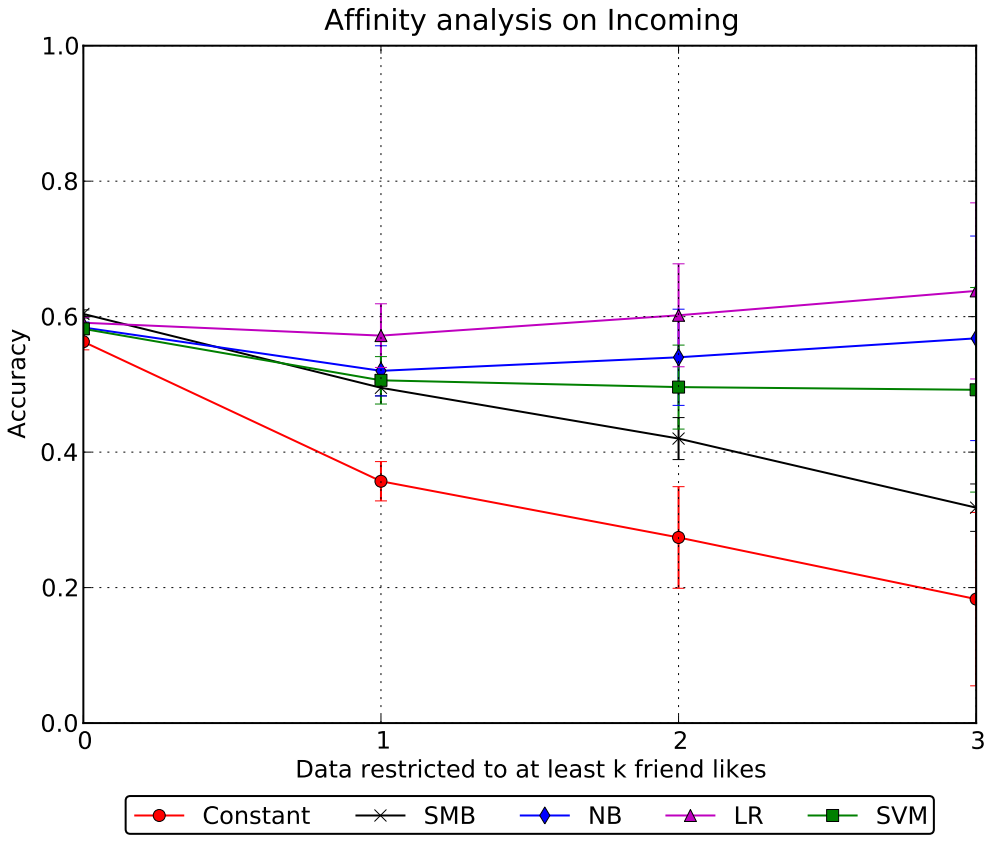


**Figure 3.9**: Accuracy results for an exposure curve using the *incoming words* feature set

Chapter 4

# User Preferences

In this section we will discuss the effects of using user preferences as the feature set.

## 4.1 Introduction

Facebook allows users to provide a vast array of personal traits and interests on their Facebook page.

Including:

- Demograhics - age, gender, location, etc
- Group Memberships
- Personal Preferences - favourite books, favourite athletes, favourite sports, inspirational people, personal interests, etc
- Conversation Data - words sent, words received

In this section we will try to uncover which user preferences are indicitive of item likes.

## 4.2 Demographics

"Furthermore, we observe a strong effect of age on friendship preferences as well as a globally modular community structure driven by nationality, but we do not find any strong gender homophily." **[?]**

Gender breakdown in the data set:

| Male | Female | Undisclosed |
|---|---|---|
| 85 | 33 | 1 |

**Table 4.1**: Gender breakdown

There is a clear male bias in the data set.

[Backstrom et al. 2011] have shown that different genders have differing tendencies to disperse interactions across genders.

| 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

Birthday breakdown in the data set:

Birthdays are grouped in a distinct range, most users in this data set are in the age range of $18 - 30$.

Undisclosed and anomlies (100 years plus) somethng.

Location breakdown in the data set:

| **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 app development and deployment) or are undisclosed, location information in this data set will not be used.

The feature vector uesd for demographics is comprised of:

- 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 demographics feature set shows positive results, which are nearly comparable to the baseline of SMB.

The exposure curve shows a positive corelation between friend likes.

## 4.3 Traits

The traits feature set inclues:

- Activities
- Books
- Athletes

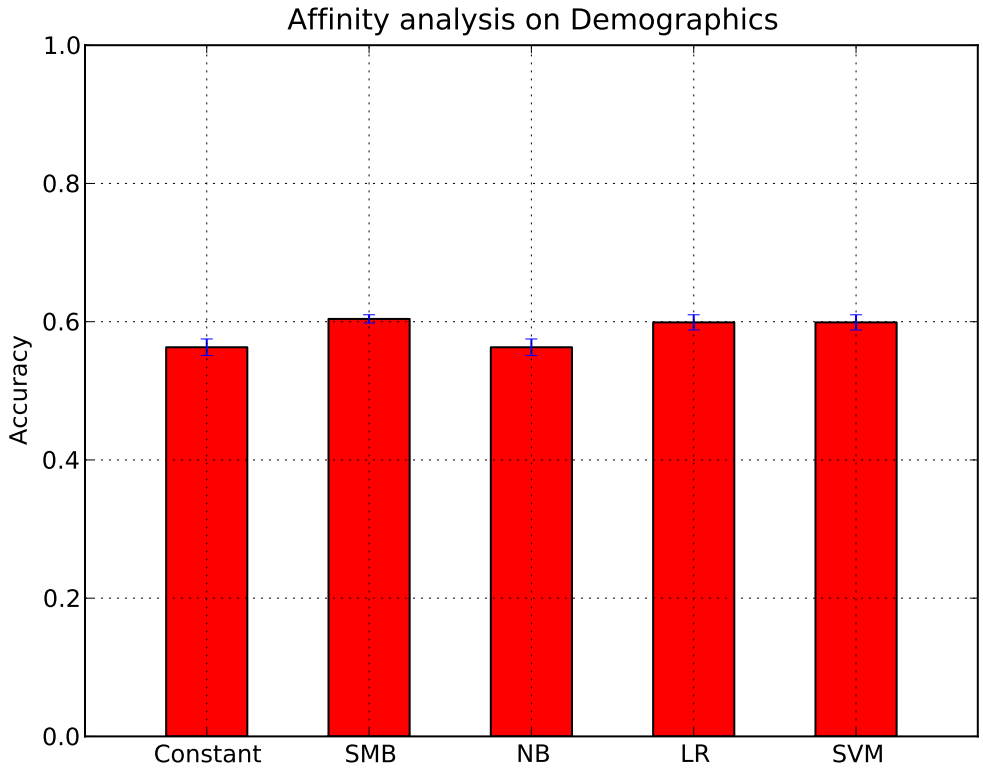


**Figure 4.1**: Accuracy results using the *demographics* feature set

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

The traits feature set shows an improvement over our SMB basline in the LR and SVM case.

This trend continues through exposure curve.

Traits show good predictive qualities the weights are:

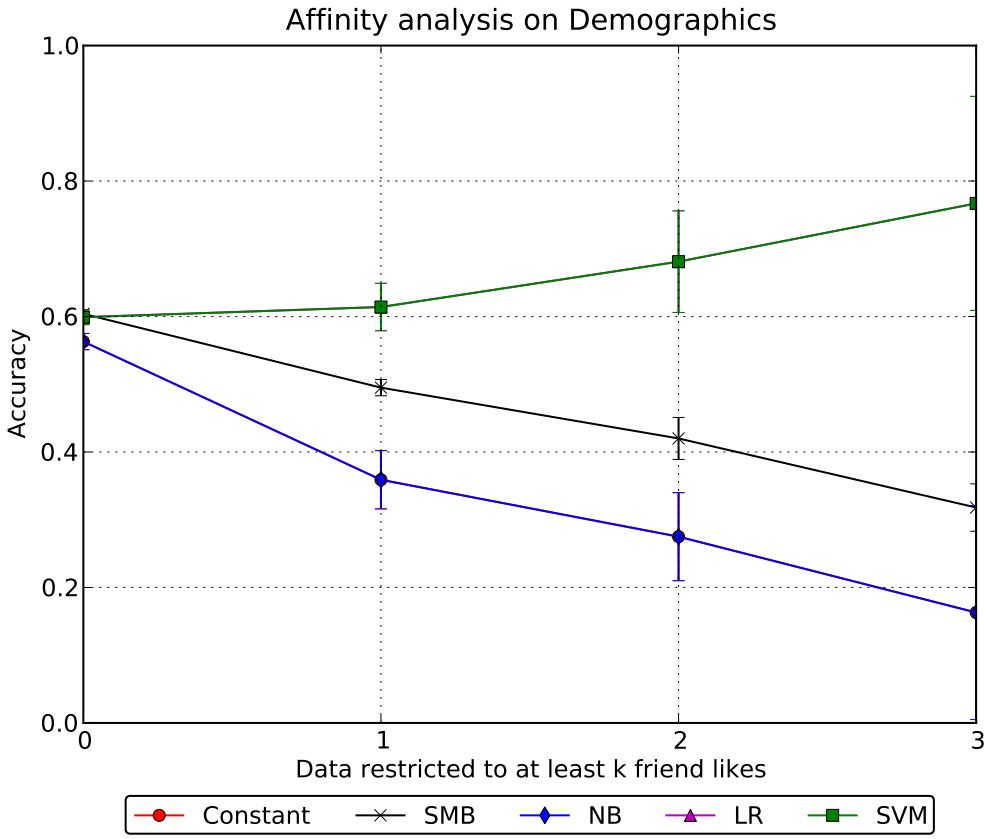


**Figure 4.2**: Accuracy results for an exposure curve using the *demographics* feature set

## 4.4 Groups

The most popular groups for app users are shown below.

| Count | 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

| Count | Inspirational Person |
|---|---|
| 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

| Count | 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

| Count | 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 4.7**: Top athletes for app users

| Count | 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 4.8**: Top interests for app users

| Count | 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 4.9**: Top music for app users

| Count | 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 4.10**: Top movies for app users

| Count | 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 4.11**: Top sports for app users

| Count | 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 4.12**: Top television shows for app users

| Count | 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

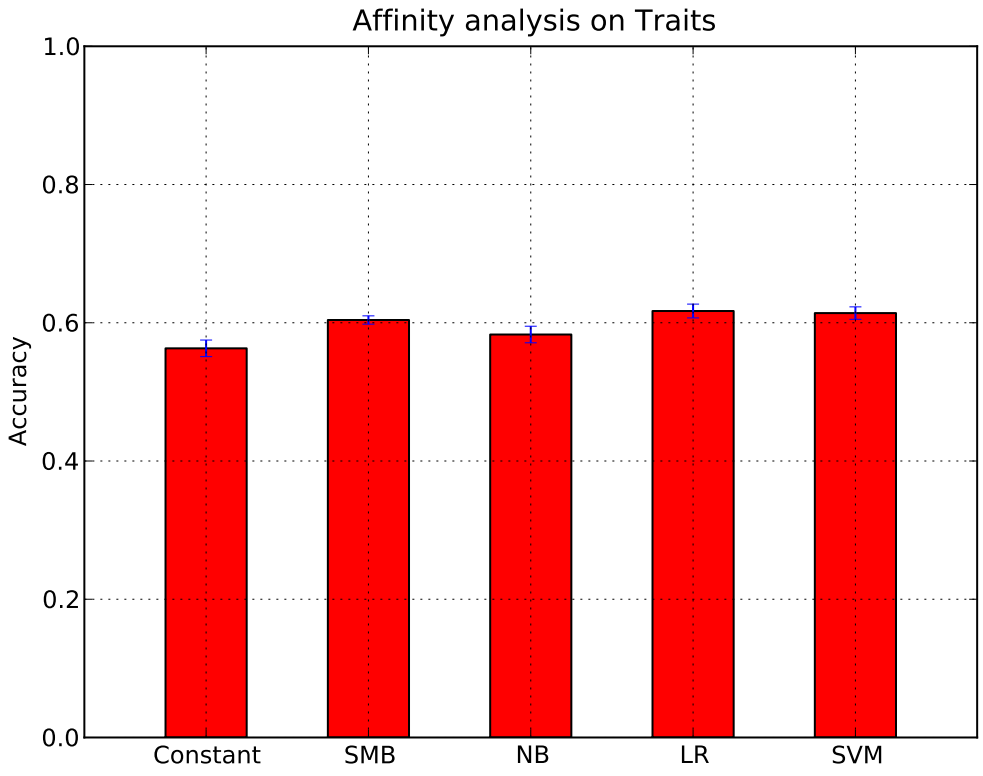


**Figure 4.3**: Accuracy results using the *user traits* feature set

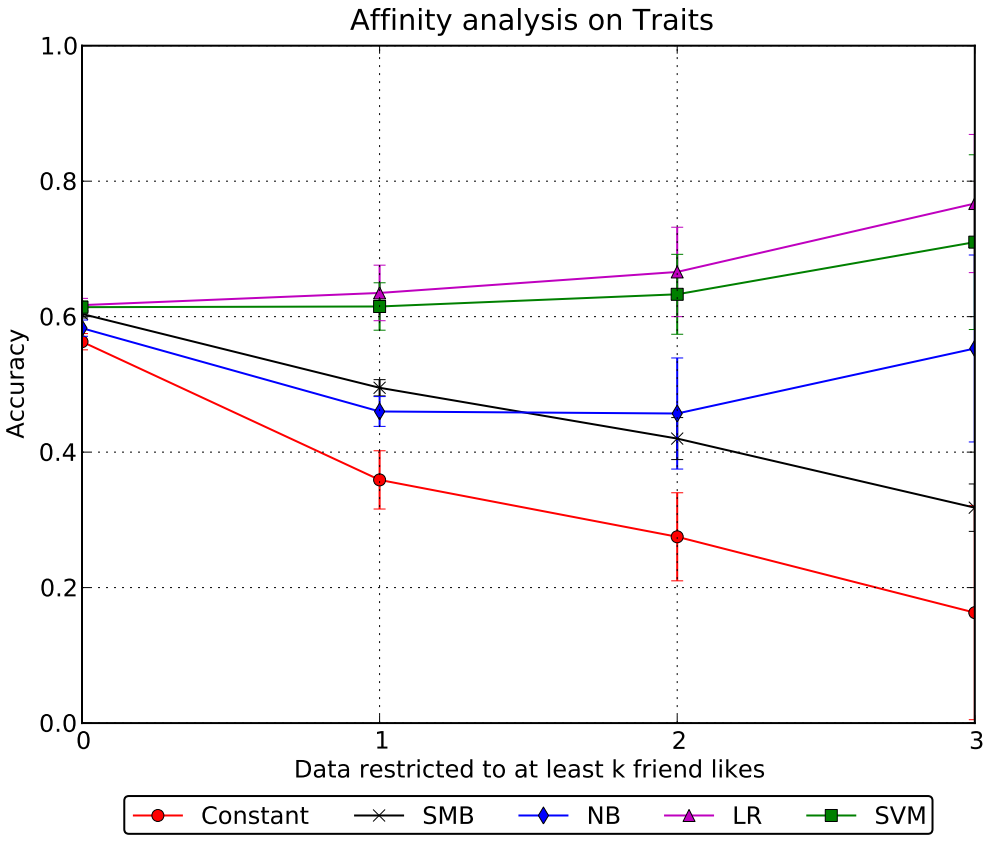


**Figure 4.4**: Accuracy results for an exposure curve using the *user traits* feature set

| **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 groups breakdown for range 5+

Given the quantity of groups on Facebook, we needed to find some optimal test size for our data set. Given memory constraints we tested in the range of 100-1000.

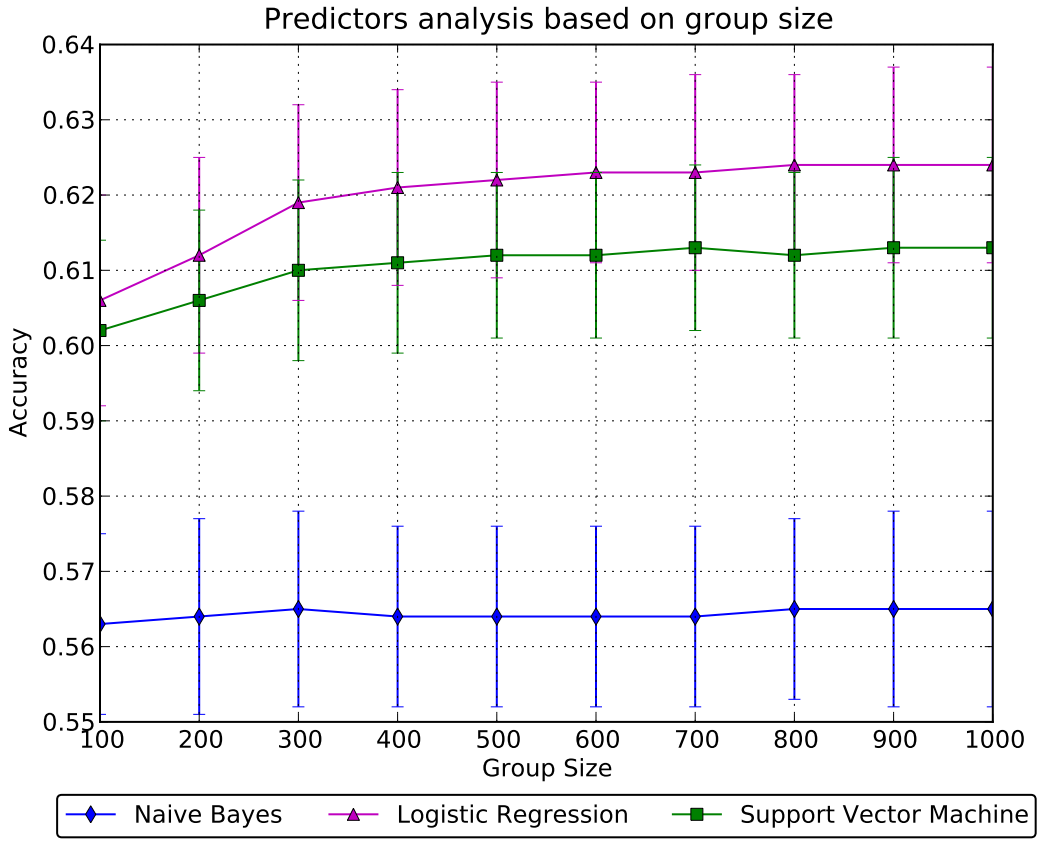


**Figure 4.5**: Accuracy results for different *groups* sizes

The most predictive groups sizes for our classifiers are:

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

Based on the analysis above LR and SVM groups outperformed our baselines.
This trend continued in the exposure curve.
Groups are predictive.

## 4.5 Pages

Page likes for app users.

The most predictive pages sizes for our classifiers are:

| 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 groups breakdown for range 5+

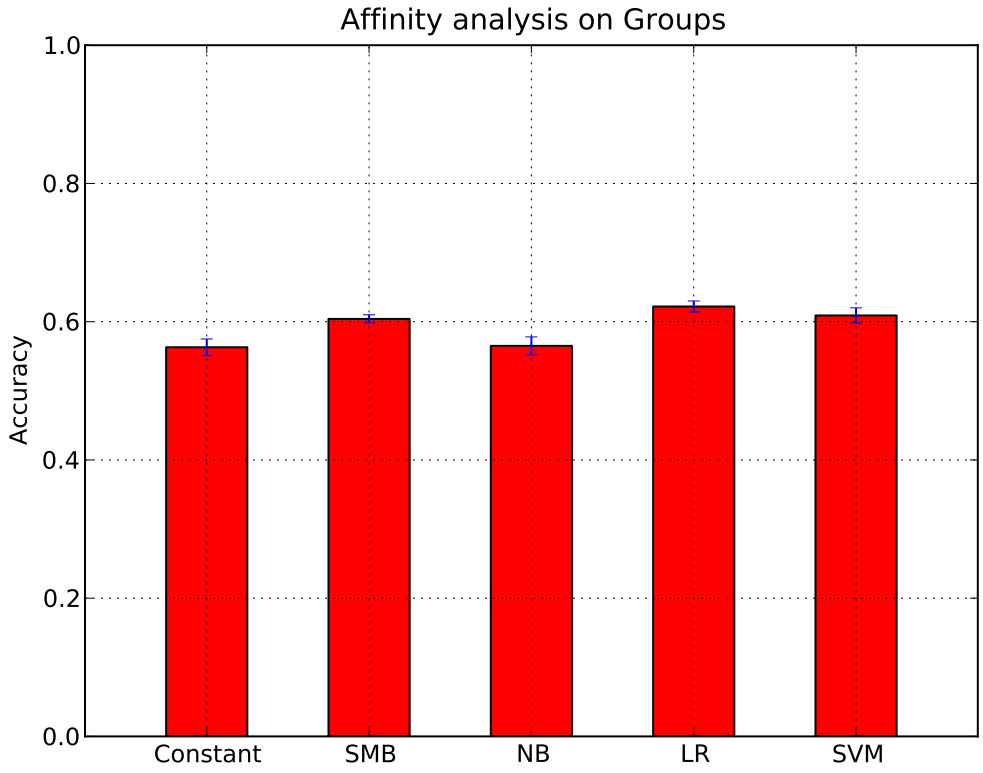


**Figure 4.6**: Accuracy results using the *groups* feature set

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

Using this optimal size we find pages are more predictive then our baselines for LR and SVM.

This trend continues through our exposure curve.

Pages are more predictive.

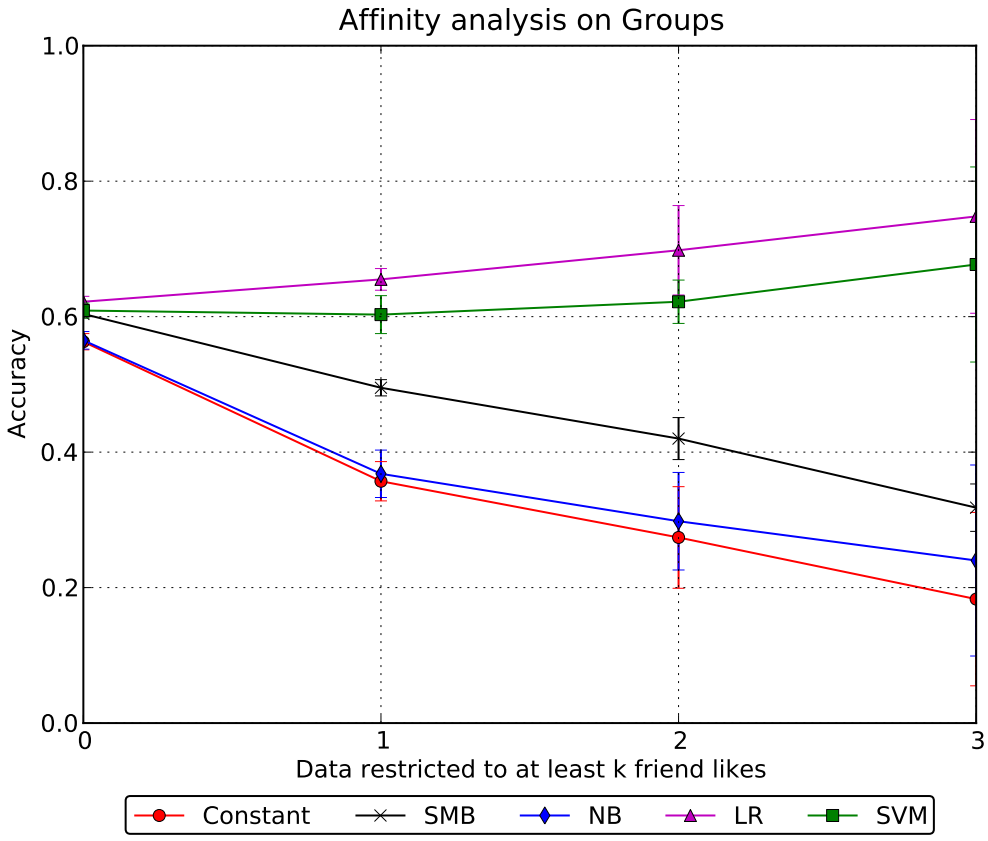


**Figure 4.7**: Accuracy results for an exposure curve using the *groups* feature set

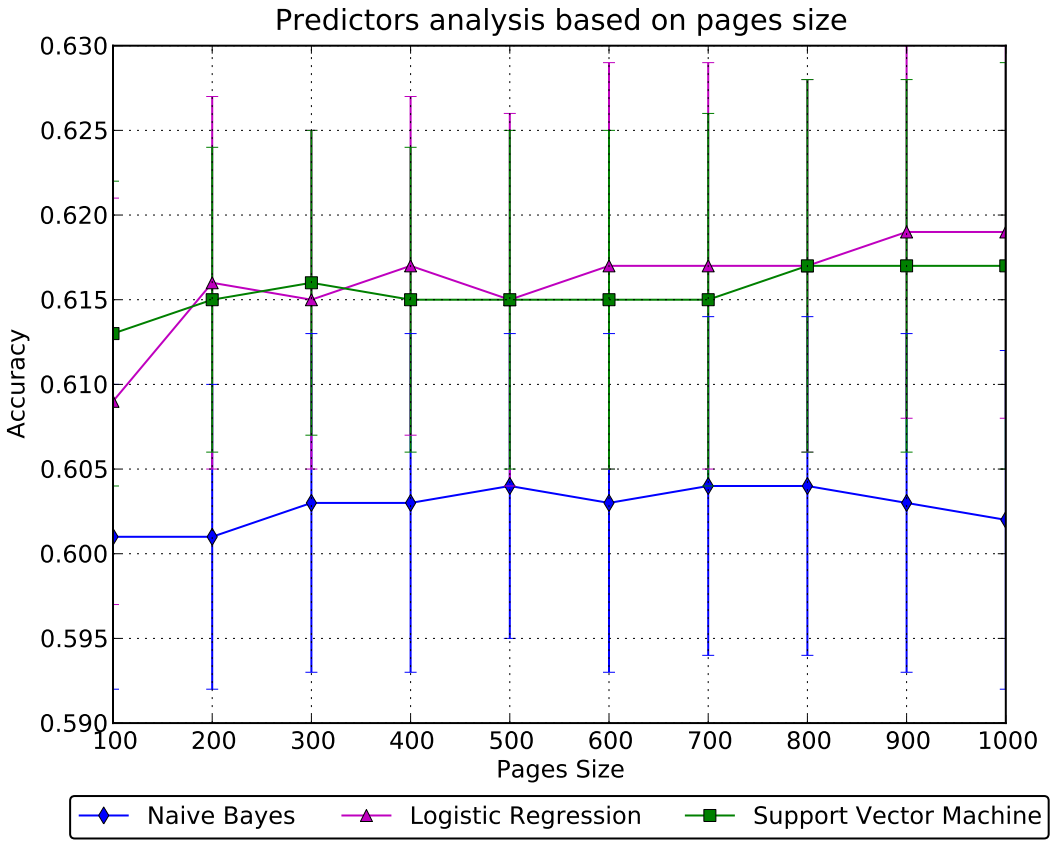


**Figure 4.8**: Accuracy results for different *pages* sizes

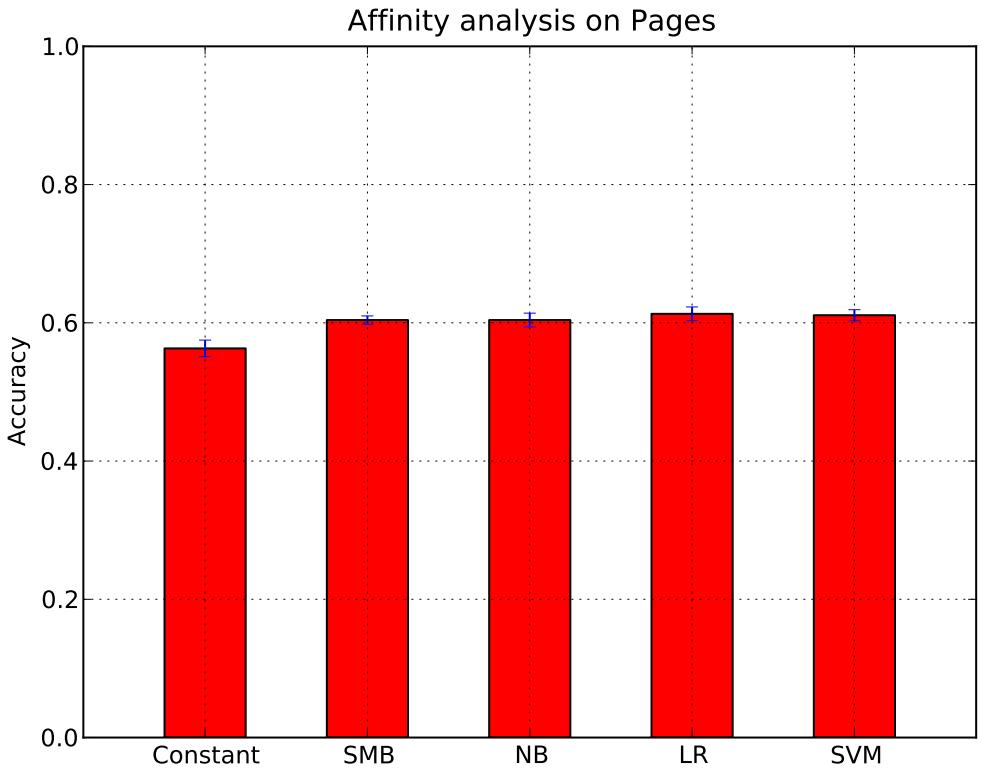


**Figure 4.9**: Accuracy results using the *pages* feature set

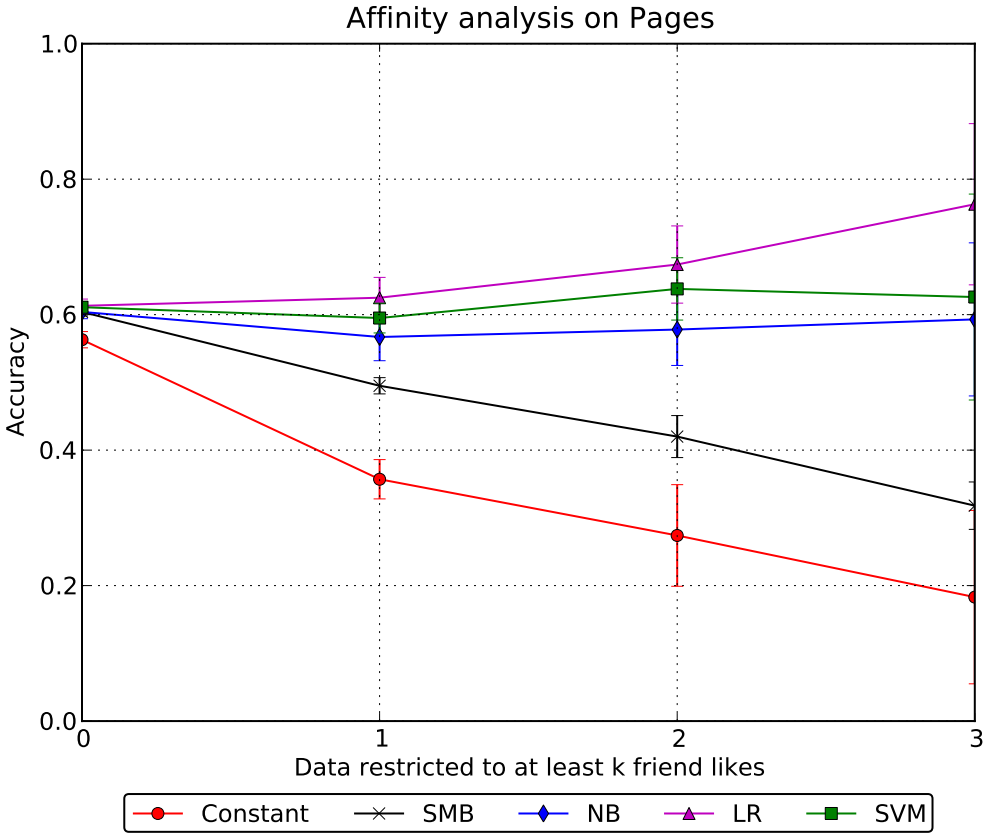


**Figure 4.10**: Accuracy results for an exposure curve using the *pages* feature set

Chapter 5

# Model Combinations

In this section we discuss two different approaches to combining the information we learned above.

## 5.1 Positive Feature Selection

Firstly, we can combine features which were successful at improving from our baseline.

Using the combined feature set of:

- Traits
- Pages
- Groups

Using these combined results we find an improvement over SMB of the form: (data here)

This trend continues over the exposure curve

## 5.2 Bayesian Model Averaging

Hopefully can implement this.

## 5.3 Summary

In this thesis we have tested different feature sets against appropriate baselines and found improvements.

## 5.4 Future Work

- Passive likes:
- Cold start:

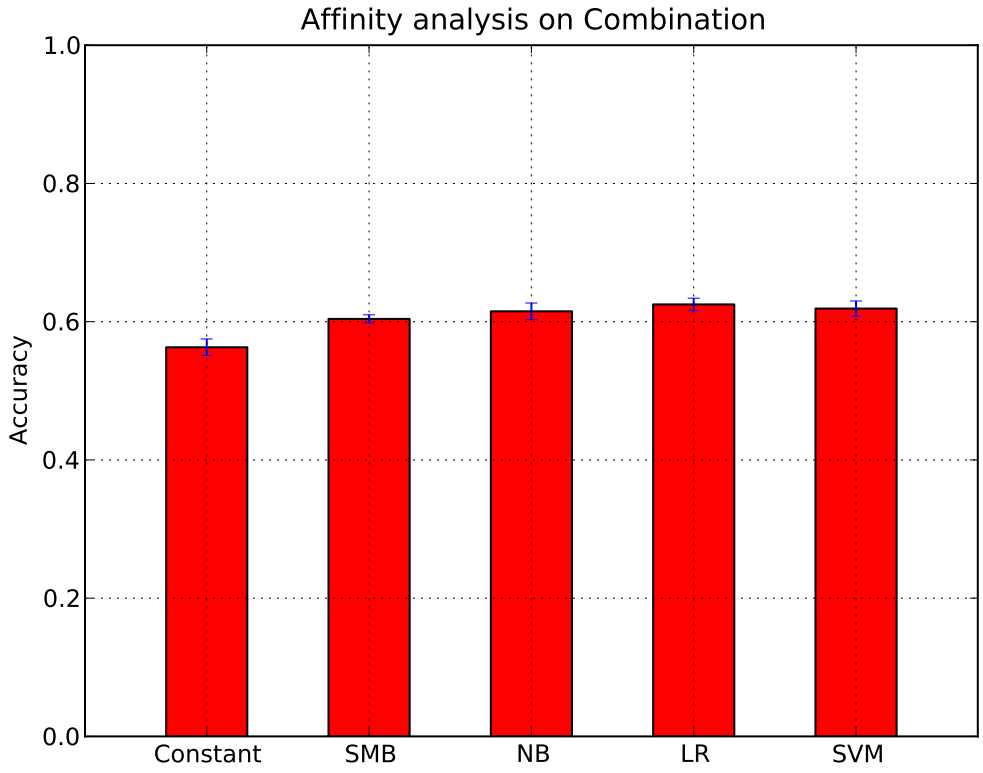


**Figure 5.1**: Accuracy results using the *combined* feature set

- General user set: Such as the study done by [**?**] which comprised of the entire active social network of 721 million users as of May 2011.

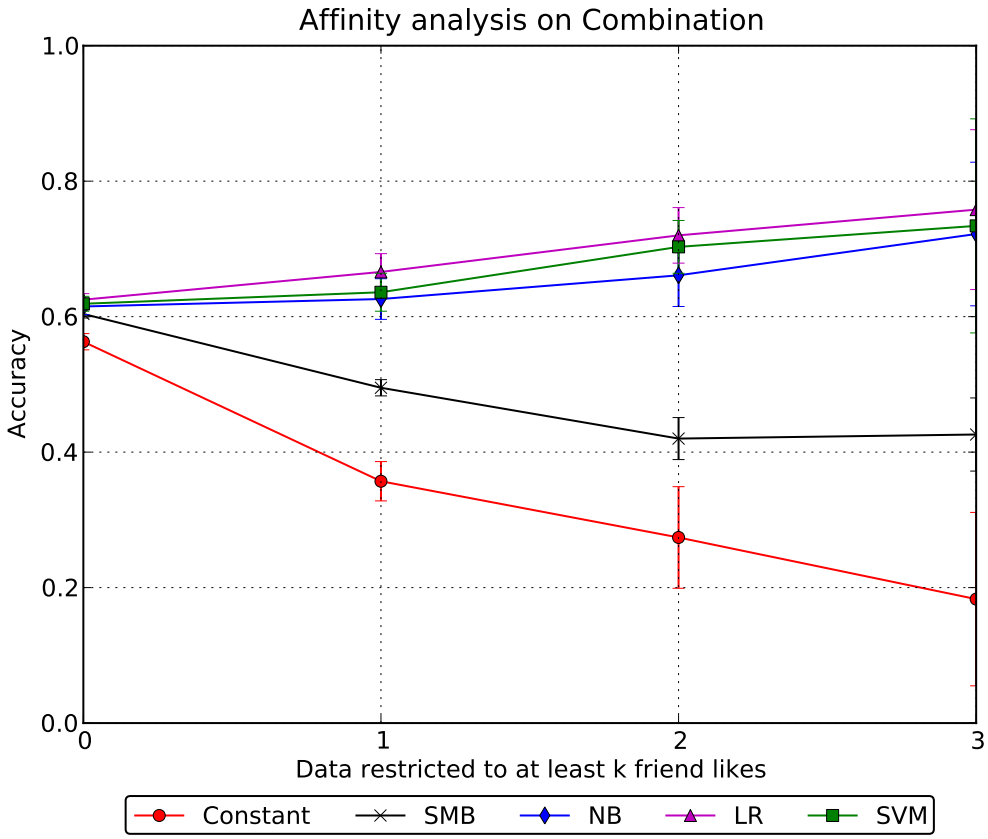


**Figure 5.2**: Accuracy results for an exposure curve using the *pages* feature set

# 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. 17)

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. 7, 23)

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*. (p. 2)

BRANDTZG, P. B. AND NOV, O. 2011. Facebook use and social capital — a longitudinal study. ICWSM'11 (2011). (p. 17)

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. 8)

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. 7)

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). (p. 8)

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)

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. 11)

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. 11)

UGANDER, B., KARRER AND MARLOW. 2011. The anatomy of the facebook social graph. *CoRR abs/1111.4503*. (pp. 23, 42)

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