PREDICTING POLITICAL DONATIONS USING DATA DRIVEN LIFESTYLE PROFILES GENERATED FROM CHARACTER N-GRAM ANALYSIS OF HETEROGENEOUS ONLINE SOURCES

by

Colin Conrad

Submitted in partial fulfillment of the
requirements for the degree of
Master of Electronic Commerce

at

Dalhousie University
Halifax, Nova Scotia
August 2015

This thesis is dedicated to anyone and everyone who believes that a great future can be made on a rocky coast near the sea.

Table of Contents

List of Tables

List of Figures

Abstract

This paper describes an approach for generating multi-dimensional Activities, Interests, and Opinions (AIO) insights from disparate web sources. The method involves identifying psychographic profiles using text analysis of social media data. The approach is tested on tweets from 438 Twitter profiles, 219 of which are integrated with filing records from the United States Federal Election Commission, 219 others were used for control. Profiles were matched using demographic criteria and analyzed using political parties and donation values as labels. Standard probabilistic, entropy and kernel based approaches are used to make predictions based on word n-grams, while the CNG technique is explored as an alternative. Using CNG two predictive models were created that were able to exceed benchmarks extracted form the literature. Using these models, we are able to demonstrate a method for generating qualitative psychographic profiles, which can in turn be used to label customers for marketing insight.

List of Abbreviations and Symbols Used

AIO	Activities, Interests, and Opinions
CNG	Common N-Gram
EII	Enterprise Information Integration
ETL	Extract, Transform, Load
FEC	Federal Election Commission
LOV	List of Values
NLP	Natural Language Processing
PAC	Political Action Committee
PIPEDA	Personal Information Protection and Electronic Document Act
PRIZM	Potential Rating Index for Zip Marketers
SRI	Stanford Research Institute
VALS	Values, Attitudes, Lifestyles

Acknowledgements

I would like to thank Naureen Ali for her patience though the countless hours we spent together working on the original ELMS implementation. Thanks go to Qigang Gao for his leadership and guidance through the NSERC Engage project. I would like to acknowledge the incredible support of people at the Dalhousie Natural Language Processing Group, especially Dijana Kosmajac who was helpful for clarifying the implementation of statistical models in NLP ... and for letting me drink her tea through the thesis writing process. Farrukh Momin deserves a big thank you for being a great friend and business partner, and for motivating me to remember to learn to code for the right reasons. Finally, I would like to offer many, many thanks to Vlado Keselj, who not only supervised this thesis, but also had to put up with entrepreneurial antics and philosophical musings.

Chapter 1

Introduction

The primary aim of a marketing effort are to effectively identify and satisfy customer needs and wants. In order to meet these goals, marketers will develop products and communication strategies to target specific groups, or market segments. Contrasted with mass marketing, targeting market segments can help firms develop more cost-effective strategies for maximizing communications or product placement effectiveness [42]. Political marketers are not fundamentally different. By tailoring strategies to interests of population subsets, firms, parties or PAC's can target specific consumer interests to facilitate better political positioning, policy positions and effective communications.

Market segmentation by lifestyle, attitudes and values is often referred to as psychographic segmentation. Using psychographic segmentation, prospects are classified according to segments based on the bundle of character traits, interests or behaviours they exhibit. Using a set of psychographic profiles, customers are categorized according to empirically validated entities that consist of individuals with similar interests or opinions. By segmenting this way, marketers can develop product insight through better understanding consumer motivations, which in turn aids the future development of product value.

There are currently a number of psychographic segmentation tools available to market researchers. VALS ("Values, Attitudes, LifeStyles") is a popular and comprehensive segmentation tool originally proposed by the Stanford Research Institute (SRI) International used to perform detailed market research based on lifestyle and demographics [37]. PRIZM is a segmentation tool provided by Claritas (licensed to Environics Analytics in Canada) to perform lifestyle segmentation according to postal code [3]. PRIZM and VALS are the result of considerable data-driven empirical research, which are closely guarded trade secrets [10]. Esri combines data from a number of sources to provide demographic and behavioral segmentation tools that

provide comprehensive marketing by geographic location using Tapestry^{TM} though its ArcGIS system [24]. SAS and RapidMiner provides proprietary methods for identifying segments using demographic clustering and purchase records [15, 34].

However, there is a theoretical gap when generating psychographic segments from non-traditional sources. Comprehensive psychographic profiles are generated by domain experts running through comprehensive qualitative study, while data-driven alternatives seem to utilize unsupervised learning to segment customers according to features. It is useful for marketers to identify lifestyle segments using public sources such as Twitter, which do not come with a ready-made psychographic taxonomy and require domain expert judgment. Similar public databases are abundant, and could be used to generate useful marketing insights at a fraction of the cost. It is desirable to develop a framework for identifying, testing and integrating profiles from across data sources of various qualities.

1.1 Research Question

The objective of this research is to identify a data-driven methodology for building validated psychographic profiles and performing lifestyle segmentation without the use of focus groups. This is motivated by a lack of a methodology specifically for psychographic profiling using web and social data. As the demand for web applications increases and the use of online databases for marketing research becomes more prevalent, so too does the need to identify new methods for generating marketing insights using online data. My research question can thus be formulated as follows:

Q Can meaningful or useful lifestyle segments be generated using consumer or web behavior records?

This research lies firmly in the field of electronic commerce. It combines comprehensive research in consumer behavior and marketing research with the application of data mining, data warehousing and artificial intelligence to business issues. We apply our research to a specific case concerning political donations, and judge the psychographic profiling methods by their ability to make meaningful predictions about political giving. Our data utilizes publicly available records from Twitter and the United States Federal Election Commission. We construct and test the significance

of lifestyle segments generated using word and character n-gram analysis. In this case, we use machine learning to identify n-grams characteristic of Activities, Interests and Opinions profiles generated from political donations and affiliations.

Though I utilize political giving records, the findings are not restricted to the domain of political giving. These datasets were chosen specifically because they combine three qualities of data: demographic, social and behavioral data from disparate web sources, and were publicly available. As our data comes from unrelated sources with no sure method of truly identifying the data's ground truth, prospects could only be matched based on similarities in their data features such as Name, City or Occupation. By investigating the dependency and challenges of these sources, we are able to identify a method that can be used to generate significant motivational marketing insight from disparate, loosely related data more generally. It is our hope that this can lay the groundwork for a larger project in market research using social media, as well as a new direction for data-driven lifestyle segmentation.

1.1.1 Document Structure and Research Contributions

This research makes three novel contributions to the existing literature. For simplicity these are listed below:

C1 To the best of our knowledge, this work represents the first attempt to outline a method for creating empirically validated a-priori psychographic segments using social media data. Though other attempts have been made to utilize unsupervised learning for lifestyle segmentation, this work extends the literature by performing an in-depth analysis of existing commercial techniques and applying the methodology used in these techniques to create a data-driven alternative.

C2 No other work has made use of the FEC (Federal Election Commission) filings for the purpose of dataset labeling user tweets. This thesis explores how these filings can be used to create reliable labels for political affiliation and past giving behaviours.

C3 No other work has attempted to apply CNG to solve political affiliation or donation problems. Though word unigrams have been utilized to make predictions

about political behaviours, this thesis makes novel use of Character N-Grams for the prediction of behaviours.

These contributions are made in the process of conducting an experiment that follows a typical design. Chapter 2 explores and analyzes the relevant literature supporting the experiment. Chapter 3 describes the theoretical framework and ontology. Following this, Chapter 4 describes the experiment design while Chapter 5 describes and analyzes the experiment results. In Chapter 6 we draw conclusions and describe three promising areas of future research.

Chapter 2

Background and Related Work

The literature concerning data driven psychographic profiling is broad and draws on elements form at least three core disciplines: Marketing, Web Mining and Ethics. Furthermore, it touches on multiple domains related to these fields such as Consumer Behaviour, Data Mining, Data Warehousing, and Privacy Law. This literature review is structured to incorporate the literature from the three core disciplines. It begins by exploring the original market segmentation problem, before identifying potential research methods to solve the problem. It then identifies the technical potential for solving psychographic segmentation problems on new media, followed by an exploration of the privacy implications of performing segmentation. Some effort is given to discussing the relationship between this review and the other chapters.

2.1 Marketing Considerations about Customer Segmentation

Customer segmentation is the process of dividing the market into its constituent parts according to some method [10]. Customer segmentation is an important subject in marketing, as it helps make meaningful predictions about consumer behavior and has even been described as "the key to marketing success" [71]. It is closely related to the concept of value creation, and is taught as a practical art in marketing and entrepreneurship [54]. By creating products and communication strategies for a group of customers with strong mutual interests, marketers can create maximal value, helping to drive sales. Effective segmentation also facilitates cost-effective marketing communication, as products can be effectively targeted to a specific subset of interested buyers. Without defining a segmentation method, marketers would be forced to either produce products for the mass market, or perform 1 on 1 marketing.

One of the oldest and most common methods of segmenting the market is geographic segmentation. By classifying customers according to geographic regions,

marketers are able to target customers on the basis of similar shared living experience. Geographic segmentation makes sense intuitively, as the market needs of specific regions might be well contrasted with others. Dennis Cahill uses the example of hotel chains: the peak tourist market in New England is the summer, while the peak demand in Florida is the winter [10]. The dynamics of selling automobiles is very different in low income countries, versus high.

Though geographic segmentation is cost-effective, it offers limited predictive power on its own. Geographic segmentation is not customer driven, in the sense that it only describes the motivations of customers that are dictated by their environment. Moscardo et al. evaluated the effectiveness of geographic segmentation in comparison to activity and interest as it pertains to the Australian tourist market. They found that, overall, activity and interest was a much more effective segmentation method, according to the standards of the tourism literature [29]. Though geographic segmentation appears to have the features of a good segmentation method, on its own, it lacks comprehensive explanatory power.

Demographic segmentation is another common segmentation method, and is perhaps one of the most popular. Markets might be segmented according to population features, such as age, height, gender, marital status, number of children or nationality. Like geographic segmentation, demographic segmentation can be collected relatively easily and cheaply. Demographic segmentation also contains significant predictive power. Gupta and Chintagunta were able to demonstrate the effectiveness of using household size, household age and income to segment populations to maximize profitability using a Bayesian model [30]. Adding demographic information to the customer segments added significant predictive power for predicting purchases of Heinz products.

However, like geographic variables, demographic variables on their own offer limited insight into why customers might wish to purchase products and the details of their behavior. Broadly segmenting customers according to race, for instance, might help make meaningful predictions, but the subjectivity of ethnicity might be less significant than other predictors [56]. Cahill suggests that broadly marketing to teenagers or women could not only miss the mark, but actually be detrimental to the marketing effort [10]. Efforts to label customers according to broad categories could

alienate consumers, and do not address the actual motivations of the prospects.

2.1.1 Lifestyle Segmentation

When we look for segments beyond demographics and geography, we might consider grouping customers according to behaviors. Unlike geographic or demographic segmentation, this has the benefit of being directly related to consumers' motivations. Behavioural segments can be defined using collections of survey data, purchasing records or usage data. Using these records, marketing analysts can perform clustering tasks to discover insights into their consumer's behavior [72]. These clustering techniques are particularly helpful when establishing methods for segmenting by perceived benefits to customers or usage purposes.

However, though these methods might be good for identifying groups of customers based on motivation, they are not particularly good at describing what the customer motivations actually are. Marketers who are searching for deeper customer analysis look to outside factors and attempt to conceptualize their customers according to lifestyle: profiles of common activities, interests or opinions (often abbreviated as "AIO"). In the context of AIO, activities can be defined as "what we do", interests as "what we want," and opinions as "what we believe" [10]. Where other behavioural approaches might focus on the product being serviced, lifestyle segmentation can be seen as a method for integrating exogenous factors into the customer segments. By targeting groups with common interests or values, marketers might be able to design products that appeal to customers' specific motivations, while still casting a wide enough net to ensure scale. Where demographic approaches might infer a set of behaviours from specific physical traits, lifestyle segmentation appeals to a set of behaviours determined by psychological or sociological traits.

Lifestyle segmentation is not new. This practice was developed by marketers in the mid-1960's, when marketers began identifying consumers according to their personality characteristics [44]. By the 1970's and 1980's, marketers began looking to segmentation based on AIO [73]. This ultimately culminated in the development of lifestyle segmentation approaches based on psychographics or sociodemographics. These are discussed in turn.

2.1.2 Psychographics

The term "psychographics" was coined by combining "psychology" and "demographics". Psychographic profiles typically refer to a series of lifestyle segments that are categorized according to a motivation or lifestyle scheme closely related to a consumer's self-concept or self-image. These schemes can be broad, as in the case of the comprehensive "list of values" tools such as SRI's VALS or Khale's public domain LOV ("List of Values") . These could also consist of industry-specific segments created to tackle unique marketing problems.

VALS is a validated framework to explain the psychological drivers of consumer behavior. Originally the brainchild of futurist Arnold Mitchell, VALS was launched by SRI International in 1978 [36]. Arnold envisioned a marketing tool that explains the relationship between psychology and consumer behavior, in an effort to develop a customer segmentation method grounded in actual customer motivations. In his book The Nine American Lifestyles Mitchell outlined nine lifestyle segments that defined the American market [48]. This lifestyle schema, though once relevant, gave way to a new VALS framework in 1989, which has become one of the most widely adopted segmentation methodologies used in the United States.

The updated VALS test requires customers to take a survey of 34 questions. Though a tightly guarded trade secret, it is assumed that these questions are the result of rigorous empirical validation. From these questions, test takers are assigned one of nine customer segments: Innovators, Thinkers, Believers, Achievers, Strivers, Experiencers, Makers or Survivors. Each segment is determined by a function of resources available to the consumer, and their primary motivation. Innovators, with high resources and high innovation cater to upscale and innovative products, while Survivors tend to purchase discounted merchandise. Other segments are paired based on primary motivation and resources. Thinkers and Believers are purported to both be motivated by ideals, though the greater access to resources makes thinkers value durability and quality over brand loyalty. Likewise, Achievers and Strivers are similarly motivated by achievement, while Experiencers and Makers are motivated by self-expression [36]. Figure 2.1 below is provided by SBI to summarize the current psychographic profiling method.

The validity of the VALS framework, though widely adopted by industry, suffers

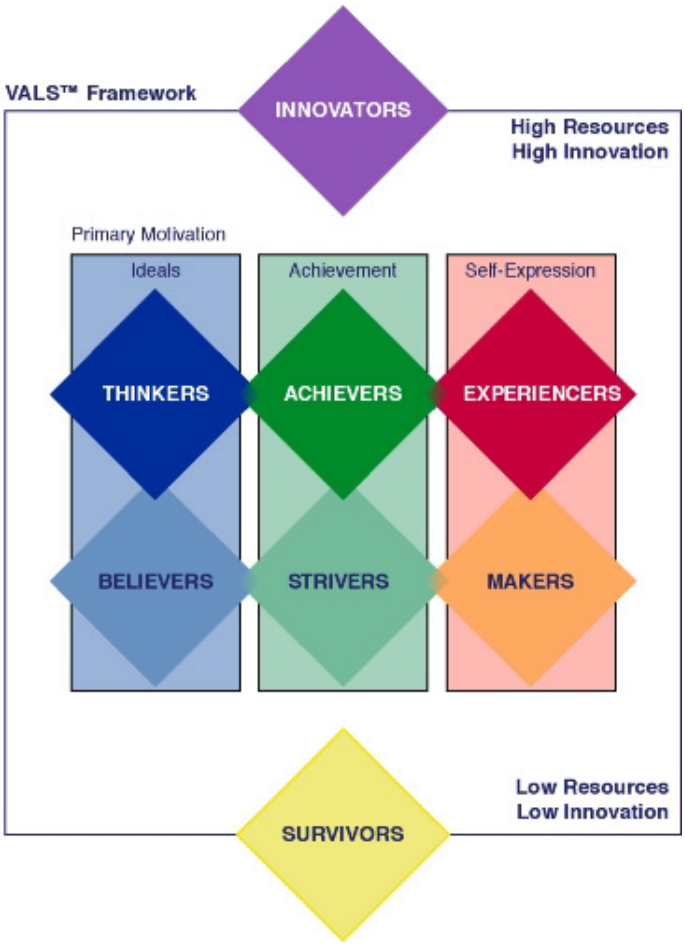

Figure 2.1: The VALS Framework [37]

from a lack of transparency. Though we can assume that VALS has been statistically validated, the rigor of the psychographic framework cannot be critically assessed. This issue led Lynn Khale to develop the List of Values (LOV) as a transparent alternative to VALS [41]. To construct LOV, Khale et al formed a series of hypothesized survey questions based on his experience with Maslow's hierarchy of needs and similar work towards value fulfillment. LOV was then validated through a series of intensive probability studies of 2264 Americans on an extended face-to-face survey [40]. Khale et al. investigated the statistical significance of variance by test answers to questions of customer values. The results lended support to the methodology, while later studies into the contrast with VALS determined significantly better R2 for LOV at the prediction to the answers to survey questions related to consumer behavior [41].

Even with this empirical validation, there is still significant criticism of the way

psychographic segmentation is used in marketing. Currently, psychographic modeling is used to identify customers who might associate with brand identity. Geraldine Fennell performed a study of 20 000 respondents comparing demographic and psychographic variables for determining product affinity and brand affinity [28]. The survey found that psychographic factors played a significant role in improvement of model fit for self-concept, attitudes or opinions when determining product use and preference, but no significant increase when determining brand preference. There was overwhelming evidence against using this segmentation for brand identification, from which we can conclude that the primary use for psychographic research is misguided.

Fennell is not alone in making this criticism. Fennell's discovery continues to find support in current literature. A recent study of 171 female college students found that the LOV measure of psychological closeness to a fashion blog was not a significant determinant purchase intentions [31]. In a Harvard Business Review article, social scientist Daniel Yankelovich and David Meer criticize psychographic schemas as being weak, when compared to models build from transaction records [81]. Instead, they recommend that firms perform alternative non-demographic segmentation using data features. By segmenting households according to purchase records, firms could then reverse develop insight into their customer's motivations using qualitative research.

Though Fennell and Yankelovich appear to be correct about the misuse of psychographic schemes for producing insight into brand affiliation, psychographic models such as LOV continue to have significant explanatory power for measuring customer motivations. Fennell found that psychographic models were significant at predicting product use, while Cahill asserts that the widespread adoption of VALS lends support to its functional merit [10]. In addition, a more recent study used the LOV to measure changes in values as consumers aged [38]. In this study, hypotheses about value stability in aging were tested against data collected from a comprehensive 9 year panel study of middle aged adults. Using answers from the LOV, significant correlations were found in values as the subjects aged. From this, we are able to find some empirical value in LOV for predicting customer values.

In addition, VALS continues to be the subject of significant marketing research. For instance, Valentine and Powers applied VALS segmentation to derive insights

about the media habits of the Generation Y population [67]. This demographic segment is particularly intriguing to marketers as they are generally not receptive to traditional marketing. Using a sample of survey responses from 121 individuals from that age segment, they were able to categorize most members of this age group as Experiencers, Strivers or Achievers. Using a one way ANOVA, significant differences in media habits were revealed between these three segments. What we can conclude is that VALS does have some predictive power, but it is related to the product preferences and habits of consumers, rather than brand affinity.

2.1.3 Alternative Methods for Lifestyle Segmentation

Psychographic classification schemes such as VALS and LOV might have been widely adopted, but a number of alternative methods for lifestyle segmentation have also received use, especially in recent years. Among the first alternatives to the Values approaches is the geodemographic segmentation used in PRIZM (Potential Rating Index for Zip Markets). PRIZM assumes that customers' geographic location is a significant indicator of lifestyle, and segments customers into AIO profiles based on their location. PRIZM is supported by ongoing validation surveys of customer values. Though the precise validation methods are trade secrets, PRIZM has been independently validated and continues to receive widespread validation through its use in industry.

However PRIZM, like other schemes, has its weaknesses. Englis and Solomon, for instance, find that though the PRIZM geodemographic segmentation might not be optimal for identifying consumers according to their aspirational identity, but probably accurate for identifying customers by their avoidance groups [22]. This finding makes intuitive sense. Though a married 30-something entrepreneur with an income of 150 000 and a 20-something sociology student with an income of 20 000 might share the same urban ZIP, they are likely defined better by the fact that they don't have much in common with suburban dwellers, rather than sharing the same aspirational goals.

Modern geodemographic segmentation using GIS is able to get around this problem by creating geodemographic segments that are more specific than ZIP. Esri's Tapestry^{TM} Segmentation tool includes 67 distinct market segments and 14 summary

groups. Rather than categorizing entire ZIP's according to segments, Tapestry™ aggregates specific GIS data to provide a list of segments according to the ZIP or geographic region. These segments are integrated with income, age and other demographic data to give a comprehensive GIS solution [25].

Esri's Tapestry™ Segmentation was originally created using a K-means clustering approach and Ward's hierarchical minimum-variance method for grouping clusters [23]. More recent iterations created using domain knowledge and proprietary "robust" data mining techniques on datasets that include comprehensive demographic data such as age, sex, education, household income and value. The profiles were then validated ex post against comprehensive surveys that include data from nearly 6 000 products and service brands in 550 categories. This validation was used to identify the AIO of the consumer segments for the 2014 Tapestry segmentation release.

This suggests that it is possible to identify alternative methods for psychographic mining. Though we cannot compare the results of the Esri system with those of traditional psychographic schemes, we can recognize the potential for creating predictive psychographic profiles by hybridizing demographic clustering with consumer survey data. It might be that highly predictive lifestyle segmentation is possible by looking beyond comprehensive schemes and tailoring solutions to the unique data available to the researcher.

Support for this thinking goes back at least to the 1990's. Mitchell performed a comprehensive analysis of factor, cluster and discriminant analysis in the case of psychographic segmentation [49]. Using risk perception data collected from 195 UK local authority officers, Mitchell attempted to discover psychographic profiles based on risk perception that helped determine purchase of professional services. Factor analysis can be undertaken when data samples meet the Kaiser-Meyer-Olkin measure of sampling accuracy, and using could be used to test hypotheses against user responses to risk variables. By finding significant factors, profiles could be constructed. Likewise, using k-means procedure for clustering on the risk-factor survey data was found to be an effective method once k was established by industry expertise. In Mitchell's case, he was able to compare a three cluster with a six-cluster approach, finding a more favorable τ coefficient to three risk clusters. Mitchell also found utility in multiple discriminant analysis.

Following the statistical evaluation, Mitchell set out to find the most salient characteristics which separate each cluster or segment. In his case, using demographic data to validate the segments, he was able to classify the segments into "Young recruits," defined largely by the highest financial risk and low psychological risk, "New timids," with the highest overall risk, and "Old-timers," with the lowest overall risk [50]. This results in a method for defining statistically significant psychographic segments that did not utilize demographic content in their formulation. However, Mitchell cautions that even though they are statistically significant and unique to the sample data, incorporating demographic data is essential to making the data useful to marketers.

If we know that lifestyle segmentation is but one tool in the effective market segmentation strategy, the question must be raised about when lifestyle segmentation adds unique value. One prominent example that receives considerable empirical support is in predicting "ecological" products such as organic food or environmentally friendly wines. A recent study found that a list of psychographic values concerning benevolence, conformity or power was significant for determining a consumer's willingness to pay for environmentally friendly wines, though was not significant for predicting actual purchases [51]. This helped the authors conclude that green marketing should primarily target the tangible benefits of the product over the intangible benefits to the environment.

When applying lifestyle segmentation to explore local and organic food consumption, Nie and Zepeda used a collection of survey data collected from 956 US food shoppers in 2003 [53]. Variables in the lifestyle sample included discrete responses to questions about freshness, taste, safety and convenience. In addition, demographic information was collected. The data was clustered using k-means initially using a random number of clusters, evaluated by distance measures. It was found that four lifestyle segments were identified, driven largely by the lifestyle data. The authors referred to these segments as "rational," "adventurous," "careless," and "conservative uninvolved". The authors were able to conclude that based on the most common value responses, specific segments would be receptive to specific features of organic food. This method can be viewed as validation of how Mitchell's method can be applied to a specific business problem.

2.2 Prospecting Using Web Data

In the world of online web marketing, lifestyle segmentation is not always a primary concern. Industrial solutions to segmentation and prospect classification problems are often data-driven and particular to a set of problems. Recommender systems, for instance, attempt to match e-commerce users with products relevant to their tastes. Tastes are often determined using behaviours such as ratings, tags, or click streams, depending on the particular application. Though the logic that drives a recommender system is highly domain dependent, they often incorporate behavioral data, opinions and demographics to generate prospects [57]. The logics can utilize a number of data sources, such as click streams, ecommerce cart content or reviews. They can also be constructed using domain expertise, or data-driven approximation using clustering or supervised learning.

Data-driven lifestyle market segmentation is not fundamentally different from recommender systems, in the sense that they both utilize customer data to deliver predictive marketing. The difference is that recommender systems often make predictions to an individual consumer, rather than about a group of consumers. While lifestyle groups can be understood as collections of individuals that follow a pattern of activities, interests or opinions, it suggests that there is room to bridge the literature on recommender systems to the lifestyle segmentation literature produce a method for producing lifestyle segments and marketing research from web data.

If the goal is to generate a superior data-driven solution to psychographic profiling using our sources, we need to ensure the sources have rich AIO features. We have procured data from three sources which contain demographic features, donations records and Twitter profiles. One approach to this is to use supervised learning. After integrating the data, a domain expert can evaluate the profiles manually, creating labels, which can be replicated by an algorithm [1]. Using this method, varieties of lifestyle group profiles can be generated with various levels of specificity, and later tested for relevance by algorithm classification performance.

This method require experts who are willing to label the data. Seeing that finding domain experts in donations marketing will be difficult, a requirement of this research is that the classifier system not utilize expert-generated profiles. Prospective applications of this research might generate expertise by utilizing customer reviews

or behavioural data. However, this might be met with resistance as prospecting firms currently follow their own proprietary profiling requirements, and might not desire to share their trade secrets to develop products that might benefit competitor firms.

Many recommender systems used in e-commerce follow a semi-supervised approach, where products are recommended to users based on demographic generalizations generated through unsupervised machine learning on a large set. However, these same recommenders use clickstreams and customer ratings to later generate recommendations based on their specific interests and requirements [82]. The recommender system specified by Zanker and Jessenitschnig used customer ratings and clickstreams to generate labels for products. Using this method, a system for rating prospects could be generated without the direct engagement of the end users.

Other recommender systems utilize only unsupervised learning to identify prospects. Weng and Liu (2003) describe a system for lead generation by clustering customers by interest, rather than the prospects themselves [75]. Customer profiles can be generated according to previous purchases. These purchases are then clustered, and recommendations are made based on the nearest neighbors to the cluster identified. This method could be applied to the prospecting problem when prospects are clustered according to the feature to be matched. For instance, prospects could be scored based on the users' past preference for high-wealth individuals who made few but large donations. Similar techniques have been successfully applied to different problems, using different ontologies [74].

It is important to note that regardless of the recommendation technique, prospect identification will be underdetermined by data matching and integration success. Optimal scoring analysis is performed when data is integrated in a way that multidimensional analysis is possible. Weng et al. (2009) demonstrated the success of using integrated data with an Online Analytical Processing (OLAP) model over recommendations generated from disparate sources [76]. We can conclude that performing analysis using features from an integrated dataset is preferred over analysis done on individual sources.

2.2.1 Record Matching

One of the chief challenges of performing prospect generation on web data is that the sources are disparate and heterogeneous. Unless users manually opt to provide keys for integrating data from across sources, a system must be developed that can identify quality matches from the various systems. Record matching is trivial for cases where same prospects share the same identifier across or within datasets. However, our project involves using datasets from different data providers with no built-in mechanism for system integration, we thus have no unique identifiers shared by all datasets.

Data integration from disparate sources is not a new problem. Data Warehouses, which are standard tools for performing analytical functions in business, integrate data from disparate sources and provide support for analytical activities [58]. The "extract,transform, load" (ETL) functions of a data warehouse follow standards for extracting and transforming disparate data for use in business analytics. Enterprise Information Integration (EII) technologies are often used to provide integration solutions [61]. In the early days of data integration, companies custom-coded their integration solutions [63]. Today, most industry data warehouses integrate using a built-in EII and data capture that comes standard with the data warehouse package.

However, there is often a need to build a specific matching solution that integrates data using non-standard features. In our case, the disparate databases contain only names and address records, which are not suitable to matching standard solutions. Data warehousing of disparate web data has caught considerable attention recently, and the efficient processing of federated dynamic databases is a live question [8]. Though there have been other attempts to develop ETL from disparate web sources, it is not clear whether tuple-by-tuple record matching has been performed to identify matches on highly sparse records or on name record alone. As such, there is need to identify a novel approach to our data integration problem.

We identified two broad approaches to record matching. One approach to do exact matching on identifiers such as names, address, and date of birth. These are sometimes called deterministic or rule-based techniques. Exact matching requires standardization or cleaning of identifiers with variable formats, such as names or address. An issue with standardization is that it does not cater nicknames (e.g. "Bill"

for "William") without the use of a comprehensive dictionary or lexicon. To address this, we could instead use phonetic algorithms such as Soundex, or the Levenshtein distance [52] to capture a large amount of the variance. The success of this approach is highly dependent upon the quality and complexity of the data. Slightly complex data will require the generation of a large number of rules in order to match records. This is not only time consuming, but it requires constant maintenance as the data changes [68].

Another approach is approximate matching, sometimes referred to as "probabilistic". For this approach, records are matched on the basis of similar characteristics between records using some sort of machine learning, such as Bayesian techniques, clustering or neural networks. Here wider range of identifiers are considered and assigned weight to the identifiers to reflect the ability to correctly identify a match. Probabilities are then compared and are accordingly assigned a confidence rating based on the criteria [69].

Though advanced techniques such as clustering and neural networks appear to perform well, it is clear that both conventional and data mining matching solutions have strengths and weaknesses and no one solution appears superior to the others. Conventional soundex or distance-based solutions produce similar results to clustering techniques without the effort involved in preparing the data. This is a major advantage when using web data, which regularly requires preparation. Importantly, though some approximate techniques appear to have additional features, they might come with systems performance burden or complexity. Neural networks, in particular, lack operational transparency, while Bayesian networks are slow to adapt to new situations [79]. This is concerning for cases of web data mining on larger databases, as running complex algorithms would prove computationally expensive.

An ideal candidate for matching might be using a trained matching tree solution. Dey et al. propose a matching tree solution that performs similarly to other high-performance linkage solutions while maintaining communication efficiency [19]. This solution is similar to a decision tree induction approach, which requires supervised learning and a reliable method for performing said supervision. Though this technique is promising, it exposes an operational limitation on experiments with web data. In order to perform such tests, we would require substantial training data, which we

might not be able to obtain without the use of a large number of data labels, such as those generated by users by manually matching.

Though traditional probabilistic techniques possess many of the features of machine learning techniques, they might perform poorly when compared to more advanced formulas. Wilson compared a number of different probabilistic techniques and recognized that they each perform no better than the Naïve Bayes classifier [77]. Wilson demonstrated the superior performance of neural networks, which according to Wang [69], contains the same classification performance domain as rule-based approaches.

However, Wilson's dataset consisted of 80 000 homogeneous genealogical records, each containing similar fields and data content. In our case, the data is truly heterogeneous, as the only fields each dataset is guaranteed to share is surname, given name, given name and profession. Even though records might share address fields, there is a high probability that these might contain different strings though belong to the same person. We hypothesize that, given the lack of common fields, we will attain minimal benefit from using advanced techniques from neural networks or supervised learning. Bilenko et al. might invalidate this supposition, as they found modest improvements by adopting clustering solution to online shopping record matching problem, though could not optimize the solution to maximize precision [46]. These problems are similar to ours, and lead us to conclude that certain machine learning techniques might yield modest improvements to the raw matching approaches, though they might not maximize precision, which is the most important metric for lead generation.

2.2.2 Identifying Lifestyle Segments from Social Media

Migueis et al. describe a method for performing psychographic profiling using company transaction records collected from a retail card program [66]. Similar to the recommender system described in Weng and Liu (2003) [75], Migueis et al. clustered the customers according to their purchase records. Given that there were a wide range of purchase records representative of the purchasing patterns identified by different psychographic schemes such as VALS and LOV, they were able to identify six clusters based on purchase baskets. Given the nature of the purchasing data, Migueis et al assert that these represent lifestyle segments.

However, this is not fundamentally different from the general behavioural approaches described earlier, where groups of customers are identified by a set of behaviours, rather than their deep motivations. Lifestyle segmentation, in the sense we are using it, relates internal findings to a set of broad external characteristics that define a customer's lifestyle. In order to produce the insights purported by the more popular lifestyle segmentation tools, this post-hoc clustering must be at most a component of a larger scheme.

Increasingly, web marketers are leveraging social media data from Facebook or Twitter to identify customer trends and characteristics [4]. Using techniques such as sentiment analysis [39], brand mentions and network analysis [20], researchers are able to identify user interests and produce customer insight. These data driven approaches to social media use techniques in natural language processing (NLP) to identify the sets of user interests. Using social media, we would be able to produce insights about customers' broader life interests.

Assuming one is able to integrate the social media profiles with a subset of users, there is a question about how this data could be interpreted to create a model. Cooil identifies three broad methods for performing customer segmentation: a-priori, post-hoc, and hybrid approaches [18]. Migueis et al. use a post-hoc approach to identify lifestyle segments, while tools such as VALS and LOV use an a-priori approach where data is interpreted against a predefined model. If we were to utilize Twitter data, it would make sense to construct a-priori interest profiles similar to VALS and LOV, as social media users only ever represent a subset of the population [21] and because of the challenges of linking social media profiles to commercial data [12].

2.2.3 Political Opinion Mining on Twitter

Though there have not been any attempts to generate data-driven psychographic profiles specifically from Twitter, there is a considerable body of work concerning the classification of political orientation using Twitter data. Twitter has been widely regarded for its rich opinion content, and seminal work in this field [55] has established methods for opinion mining using positive/negative sentiment and a subjective/objective (neutral) axis. This approach is similar to the "sentiwordnet" sentiment analysis lexicon [26] publicly available for sentiment analysis performance.

Early attempts at identifying political opinion using sentiment analysis involved mining tweets retroactively from election events, such as the 2009 German election [65], and analyzing tweets targeted at candidates. Using sentiment analysis, Twitter could be utilized to predict election results with accuracy similar to opinion polls. Conover et al. [16] utilized a manual labelling approach on 1000 Twitter users to identify profiles according to political sentiment polarity. With these labels, they were able to examine two methods: content (hashtag) and network (following) analysis, ultimately concluding that a combination of these methods produced the most accurate results.

Cohen et al. cite a number of fundamental challenges challenges with classifying Twitter profiles according to political giving this way [14]. Among these challenges is a disparity between highly active political actors and regular users. Using a dataset of self-declared political affiliations, Cohen et al. distinguished Politically Modest users from Politically Active users and Political Figures. They used Amazon Mechanical Turk (AMT) services to manually label the profiles according to perceived political affiliation. They found that there were severe limitations with building labels for non-politician users, and that existing methods of classifying political affiliation largely fail on non-politicians. Citing challenges with labelling, they conclude that many of the challenges are rooted in the incorrect classification of users.

Akoglu [2] describes a method for identifying political polarity ranking using unsupervised learning and network analysis. The advantage of this method is that it does not rely on labelled data to produce comprehensive political affiliation analysis. Using signed bipartite networks Akoglu treats political affiliation as a node classification problem that also considers the magnitude of political polarity. Using data acquired from congressional records and PolForum, Akoglu demonstrated a functional alternative to other methods that did not hinge on manual labeling. However, this approach is limited insofar as it only addresses highly active political actors, and does not address the problem of regular users.

2.2.4 Character N-Gram Analysis and Author Attribution

Much like data-driven approaches to profile classification, the CNG character n-gram technique has been utilized for purposes of author attribution [43], stock prediction

[9], Alzheimer's detection [64] and other tasks that detect subtle text differences. Similar to author attribution, the detection of subtle psychological differences in political behaviours might be best detected using differences in their writing. Where other models using words make predictions based on the content specific to political interests, CNG focuses on the differences in their aggregate writing for attribution. Given that we are searching for psychological differences that affect the behaviours and AIO of specific groups, this research problem is a strong candidate for character n-grams.

Unlike other techniques which might consider the probabilistic occurrence of certain events or sentiment, CNG focuses on the relative distance of n-gram occurrences between classes. By assessing the distance of a broad range of n-grams and n-gram lengths, the CNG method is able to create a classification model that considers subtle textual features and perform better with noisy data. Given the success in author attribution, CNG might be successfully applied to problems of activity, interest or opinion mining on a dataset such as Twitter.

2.3 Protection of Privacy

Though sentiment and activity mining makes use of publicly available web data, comprehensive mining involves record linkage in order to render it useful. Even though these sources meet the definition of "publicly available data" under the Tri Council Policy Statement on the Ethical Conduct for Research Involving Humans [13] web mining involving humans is often seen as conflicting with the value of privacy, or even offending against privacy rights. After all, the goal of web mining for marketing is to make predictions about buying patterns and related human behaviours, and web mining techniques make use of personal information without explicit user consent. Two privacy related considerations can be examined. The first concerns the normative and ethical value of privacy. The second concerns the conception of privacy as a right and its legal value in the Canadian context.

2.3.1 Privacy as a Right and an Instrumental Value

Before exploring the normative value of privacy, we should first explore an important distinction between data and information. The term "data", in the basic sense, is

used to describe digitized computer codes that represent phenomena; the actual code transmitted and contained processed by digital technologies. The term "information" describes the usefulness of data, or the meaning of the codes that gives the codes value. The process described in this paper can be understood as a way to make information out of large amounts of web data by linking various disparate sources to a single individual. Broadly, this technology uses publicly available data to make information, the quality of which could once only be obtained though voluntary personal disclosure. The ethics of this is thus largely determined by the ethical value of the privacy norms that once preserved this type of information.

Philosophers typically distinguish between instrumental and intrinsic value. Goods that are instrumentally valuable are valuable insofar as they are a means to an end. Money, for instance, is typically thought as a mere means to purchasing goods or services, so is instrumentally valuable insofar as it facilitates those purchases. Goods that are intrinsically valuable, by contrast, are valuable in themselves. This category includes goods that are the ultimate goal, or telos of human action, such as happiness or justice. Whether they are envisioned as a Rawlsian social contract [59] or an inalienable natural right along the lines of John Locke [45], human rights necessarily belong to this intrinsic category of value.

Some natural rights however, also have an added instrumental value. Aristotle espoused a theory of value where some goods are both intrinsically and instrumentally valuable to the acquisition of eudemonia, or human flourishing [7]. It can be argued that Amartya Sen's notion of a capabilities approach envisions rights this way [62]. Capabilities, according to Sen, are those positive freedoms humans need to have in order to achieve the things we have reason to value. Privacy likely falls into this category of right, as it is necessary to have the capability to develop beliefs or conscience that cannot be developed in the public eye. Privacy is valuable in itself, as necessary to being a successful flourishing human, but is also instrumental to attaining goods such as justice. We have reasons to desire a society where we are free to develop beliefs or conscience, especially in a society that exhibits systemic injustice.

This is supported by the tradition of liberal rights and the original conception of privacy. 17th Century philosopher John Locke lays the foundation of privacy rights in his Second Treatise of Civil Government [45]. In this, he describes a constitutionally

restrained state, in which individual's natural rights are protected. Specifically, article 87 of the Treatise outlines a notion of natural rights to an individual's "life, liberty and possessions". Possessions, according to Locke, are the private dominion of an individual to do with as she or he pleases. Later, Locke contrasts this notion with that of a public person, whose "is vested with the power of the law". Privacy, according to Locke, is thus a concept that facilitates the execution of individual rights and is to be upheld by public persons so that these rights may be properly fulfilled—a notion later rearticulated by Samuel Warren and his legal establishment for the protection of individuals against "yellow" photography [70].

Of course, the positions of Locke and Warren are also cited as examples of how rights to privacy should be conceptualized merely as an means to the end of some other value, and not with the added value in itself. In the case of Locke, privacy could be construed as merely a means to flourishing, while Warren explicitly cites a means to good families. This notion of privacy as a means-to-end is also envisioned in the Canadian Charter of Rights and Freedoms [60]. The charter does not make explicit mention of a right to privacy, instead being interpreted to make reference to privacy through other rights. Section 2, for instance, outlines rights to freedom conscience, religion and assembly while Section 7 specifies a right to life, liberty and security of the person and 8 describes a right to protection from unreasonable search and seizure, but not privacy. Other areas of Canadian legislation support the merely instrumental value to privacy as well as evidenced in legislation such as the Privacy Act or Personal Information Protection and Electronic Document Act (PIPEDA). In these cases, though privacy is treated as the property of an individual, the justification of the possession of that property comes with constrained limits. Lisa Austin at the University of Toronto supports this interpretation, arguing that one of the weaknesses of PIPEDA case law is that it sometimes fails to adequately consider the original values PIPEDA was envisioned to defend, thus fails to legislate significant choices about privacy to consumers [5].

However, as technology develops, so can the conception of rights. With the emergence of increasingly invasive innovations such as biometric or behavioural data, privacy has become less about preserving our livelihoods from gossip or government, and more about the preservation of our biological or behavioural constitution. A right

to privacy, particularly with respect to social and biometric data quickly becomes a matter of the preservation of the very things that make us human. Losing access to our digitized genomes could translate to the commercialization of our biological destinies and even lead to the systematic genetic discrimination characteristic of science fiction. Closer to reality, the Ontario Court of Appeal's decision in Jones v Tsige recognizes a limited intrinsic value to the violation of privacy by establishing the tort of "inclusion upon seclusion" [35]. As the conception of privacy as an intrinsic right develops, research in the commercialization of biological or social technologies must be increasingly mindful of the practical ethics of around privacy online.

Ethical use of record linkage should thus be those uses that are constrained to not intrude upon seclusion. Using a basic reasonableness test we can infer that some uses for linkage are unreasonable while others are reasonable. Using linkage to determine bank information would not pass this test. Determining psychographics, by contrast, is akin to making generalizations about individuals based on their behaviour in a public forum. Psychographic mining would be ethical insofar as it uses publicly available data, where consent for use is assured such as in restaurant reviews, or microblogs.

Chapter 3

Theory and Methodology

Following this literature review, it is clear that psychographic profiling is a fluid practice, intended to utilize the best empirical tools available to better understand customers. In an age where electronic and social media are dominant, there is a clear need to create new profiling tools that utilize the emerging media to benefit society. This chapter synthesizes Natural Language Processing techniques to propose a method for solving a defined problem using Twitter data and filings from the Federal Election Commission. Methods for conducting and evaluating an experiment are identified, along with some attention given to the differences between a Natural Language Processing technique and the methods employed in VALS and LOV.

3.1 Identifying the Task

If AIO profiles are designed to assist marketers in understanding the motivations of their customers, AIO profiles generated from web data should likewise give a rich understanding of customers. In order to do this, a robust psychographic profiling tool should incorporate demographic, behavioural and social data to create profiles from a subset of the population. Using this subset, we would be able to identify the relationship between the psychographic profiles and data features. This, broadly, describes how a data driven AIO profile can work. Profiles are identified using one of the three profiling methods: behavioural, a-priori, or post-hoc approaches. The profiles are then used to identify customers from the features of a dataset. Following the initial profiling, the results of the marketing task can be used to again refine the AIO model, which in turn is used to identify new segments. Repeat ad infinitum.

Our task is to test an approach to AIO profiling that integrates social media data and behavioural records, and determine how they can be applied to generate meaningful marketing research. We combine data taken two disparate web sources: Twitter, and the Federal Election Commission (FEC). The data are matched and integrated

using a predefined matching method, and are then labelled according to features reflective of political affiliation and giving patterns; a method grounded differently from Mitchell's "Maslowesque" framework [48]. After building the profiles from a subset of the integrated data, they are tested based on predictive accuracy, according to an accuracy threshold specific to the task in question. Following the development and validation of the AIO profiles, the profiles can then be used as a validated psychographic label to predict political donations. These profiles have advantages over other methods, insofar as they rely on general AIO data to predict election affiliation or donations, rather than sentiment from political tweets. This is desirable for engaging the majority of non-politically active Twitter users. It is also desirable, because it opens a domain for using social data to generate marketing insight of a quality similar to VALS or LOV.

To actualize this, I perform character n-gram analysis on each user's tweets, treating the tweets as if it were an text detection problem. Each Twitter profile is broken down into n-grams, which are then used to perform supervised learning. Implicit in these n-grams is a relationship between the social media data and giving data, which helps to paint a profile that would otherwise be undetectable to the human eye. This is quite different from Mitchell and Khale, who use responses to survey questions to test the profiles predictive significance [48]. In the case of LOV specifically, survey questions pertaining to AIO were asked of a sample, and used to differentiate the psychographic profiles [41]. Questions with a statistically significant variance were used to specify criteria for classifying a customer under the profiles. Our technique maintains the integrity of statistically significant variance, but instead relies on the predictive usefulness, or entropy of each data feature.

Rather than build profiles based on Maslow's hierarchy of needs, we can extend this technique to generate AIO profiles based on data features pertaining to common interests that are useful to a specific task. Though we never have access to comprehensive information about prospective customers, limited and rich data is frequently available for profile generation. Using behavioural records, we can then label data according to the behaviours characteristic of our desired features. This would yield qualitatively similar results to surveys, with the notable exception that these results are generated from third person descriptions of the customer, rather than first-person

perceptions of self.

3.2 Data Sources and Integration

The data used in this project comes from two publicly available sources: Twitter and the Federal Election Commission. Though data from these sources are disparate, both refer to individuals and contain information about their lives and actions. Each source consists of a number of atomic "transactions" such as tweets or donation records. Analysis of the atomic sources must feed into summaries that facilitate psychographic analysis using machine learning environments such as Weka or R. In addition, our record labels must be contained in a single table of summaries of relevant features, organized according to prospective customers. The challenge with this is that Twitter records consist of a series of 140 character microblog posts, while FEC records consist of a series of donations receipts contained in a relational database, rather than summary features. These are very different.

To prepare the data, we first need to describe what we are hoping to accomplish. In order to produce psychographic profiles of a quality similar to VALS or LOV, we should begin with our labels. The FEC filings consist of a series of transactions given by individuals, along with their names, amount given, addresses and occupations. Each transaction has a corresponding committee id, which links it to a separate table containing the details of the committee, its candidate and party affiliation. These records must be integrated and stored on a single table to meet our criteria. To expedite this process, an industry partner was willing to provide a dump of aggregated FEC filings contained on their servers. To help narrow the Record Linkage problem, the partner constrained the dump to records from the United States state of Missouri; a state containing approximately 2% of the US population.

In order to perform record linkage, we needed access to Twitter user information that corresponds to the FEC data. A second Halifax-based industry partner identified a list of Twitter profiles from Missouri on their local servers, and was willing to provide a data sample for experimentation. These profiles contain information about users' screen name, as well as their real names and self-reported city of residence. Using the Twitter ID as a primary key, we are able to extract tweets belonging to that user. The challenge becomes finding a reliable method for linking the data.

Given that our records are held in disparate databases, there is no unique key to connect these records. In order to integrate the data, we drew on past experience performing disparate record matching, which demonstrated good matching precision using additional personal identifying information, such as names, addresses and occupations [80, 17]. By performing matching on prospects based on name, city and occupation, we could be reasonably confident in matching between the two data sources. This of course yields a very small subset of the total population between the two sources. However, collecting bulk data at the state level allowed us to extract sufficient matches between sources.

Before matching can be performed however, names and addresses need to be parsed and the data needs to be cleaned. Twitter records containing unusual characters that should be removed prior to uploading on MySQL. In order to manage the problem of duplicate matches and duplicate Twitter accounts, all Twitter records with duplicate names and addresses are removed. Using a Python program, names are parsed into first and last names. Using MySQL, tuples from each table could be matched based on similar features. In order to preserve data integrity, the comparisons are saved separately—each identified match is identified by the recorded keys and details of the corresponding tables.

3.2.1 FEC Summary and Labelling

When a match is discovered, FEC data can be summarized and recorded in relation to the prospect. Given that there could be multiple FEC records corresponding to a single user, the data must be summarized before being recorded in the single table. This raises the question of producing the most effective data summaries for analysis. Each FEC record contains data on the year of the donation, as well as the amount, the candidate and the party affiliation of the candidate given to. The table below summarizes the data contained in the FEC file records.

From these files, we are able to ascertain a number of useful features that could be used to label our data. The first of course being the year of the most recent political donation. With this, we would be able to assess the reliability and relevancy of the donations records. By also recording the party affiliation and gift amounts of the most recent political gifts, we are able to record labels for prediction.

Data Feature	Type	Description
id	TEXT	Item ID
original_donor_name	TEXT	Donor First and Last Name
donation_year	DOUBLE	Year of Donation
gift_amount	DOUBLE	Amount of the Gift
donor_city	TEXT	Donor's Home City
donor_state	TEXT	Donor's State
donor_zipcode	TEXT	Donor ZIP
recipient_committee	TEXT	Name of Recipient Committee
recipient_treasurer_name	TEXT	Name of Recipient Treasurer
committee_connected_organization	TEXT	Name of PAC or Other Org
candidate_name	TEXT	Name of the Electoral Candidate
candidate_party_affiliation	TEXT	Name of Candidate Affiliate Party

Table 3.1: Donation Records

Other useful data features could include the number of gifts, the reliability of political affiliation or the aggregate value of the gifts. However, after assessing the data, it was realized that there was a comparative lack of change among political affiliations and giving patterns. Prospects who gave large gifts to political candidates are the most likely to have made past large gifts, and so on. As such, FEC record summaries could reliably be best utilized to record most recent data. The table below explains the summarized FEC data. Note that there are other features than the "recent_party" and "is_donor" used in this experiment.

Data Feature	Type	Description
recent_year	TEXT	Year of Most Recent Donation
recent_party	TEXT	Part of Most Recent Donation
recent_candidate	TEXT	Name of Most Recent Candidate
gift_max	DOUBLE	The Value of the Most Recent Gift
gift_cycle	DOUBLE	Total Given over the most recent Four Year Cycle
gift_total	DOUBLE	Total Given
num_gifts	DOUBLE	Number of Gifts
is_donor	BLOB	Whether prospect has given
is_big	BLOB	Whether prospect has given gifts larger than $1000

Table 3.2: Psychographic Taxonomy

These summary features were created based on three considerations. The first is the assumption that the most recent data is the most reliable data. By selecting the most recent donations as the determinant of party and gifts, we are able to make

more reliable selections with the Twitter data, which is collected in the present. The second consideration is that gift cycles might be more relevant recent gifts. In the donations industry, gifts are often solicited in cycles loosely based on the life of the campaign. In this case, Presidential candidates are elected every four years. By focusing on the most recent set of four years, we might aggregate recent political activity that accurately incorporates the large presidential campaigns, even if they are not necessarily the most recent. Finally, we consider that a class that distinguishes donors and large donors from non-donors. By using these broad class features, we create a classifier with higher accuracy than if we were to focus on specific traits.

Following the integration of the FEC files, there are Twitter profiles that could be matched on the basis of name, city and city using a MySQL query. In order to ensure matching quality, each profile is manually assessed for plausible matching on the basis of occupation. Each match was determined on an evidence basis, where records were retained only if there was some evidence of an occupation match. In this sense, the data sample is sifted to select a small number of users who can be assessed reliably.

3.3 Political Profiling Using Twitter

When seeking charitable and political donations, there is not only a need to identify individuals who have political affinity, but also have high income and a willingness to give. Tools like the Donor Pyramid [6] are commonly employed by donation prospectors to identify their needs: a large amount of donors willing to lend small amounts of support, a small number of donors willing to lend large amounts of support. Recognizing that in American political affinity can be best understood as a polar, and by identifying a need to detect high-income individuals, we can distinguish four types of psychographic profiles: Liberal Supporters, Liberal Large Donors, Conservative Supporters, Conservative Large Donors.

The challenge of making this distinction is data labelling. Where Cohen et al. [14] cite a challenge with political labelling, and even resort to Mechanical Turks for classification. Using donation records from the United States Federal Election Commission we are able to not only generate labels for political orientation, but also estimate giving capacity.[1] This overcomes one of the primary challenges cited in the

[1]All individual political donations are publicly available from the FEC website at

Twitter political affiliation literature, which suggests that political donations as a particularly useful domain for a wider thesis on online psychographic profiling. By combining Twitter and FEC records, we are able to integrate activities, interests and opinions data with actual behaviours, reflected by political donation transactions.

We envision a political profiling system which accounts for two broad profiling considerations: "affiliation" and "propensity". Political affiliation is a measure of an individual's political preference or opinion. In this case, we envision a polar system between liberal and conservative preferences, similar to Akoglu [2]. Propensity, by contrast, consists of an individual's likelihood to give to political causes. Given that there are some political supports who do not give to politically affiliated causes and are yet active in their political opinion, using these two dimensions we can envision dimensions somewhat similar to those expressed in VALS. Drawing on our four profiles, we can express the profiles with the image below:

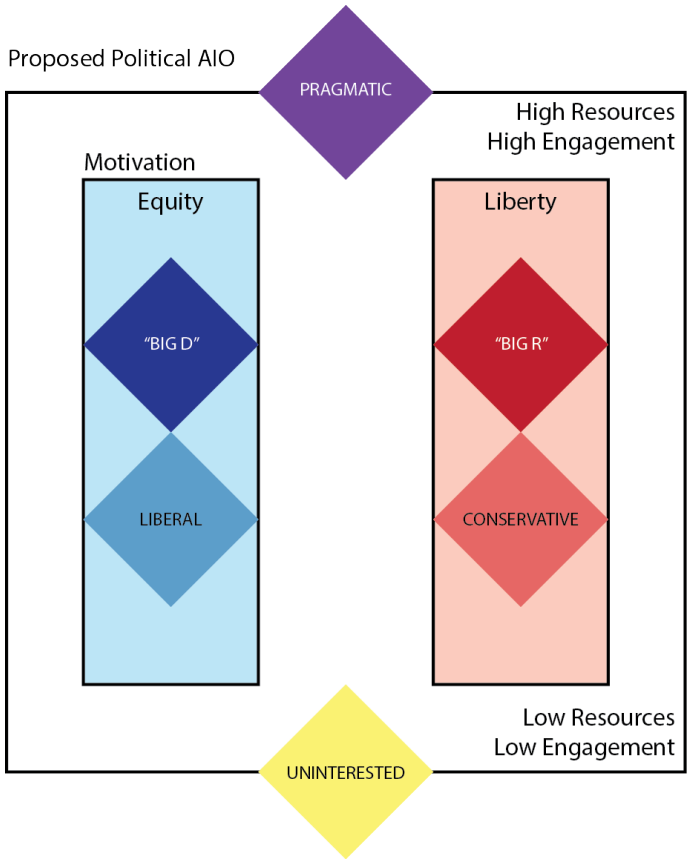

Figure 3.1: VALS-Like Imagination of AIO Framework

http://www.fec.gov/finance/disclosure/ftpdet.shtml

Like VALS these political profiles account for Activities, Interests and Opinions to distinguish profiles of political affiliation. Also like VALS, these profiles can be validated through predictive significance. However, the methodological difference between the systems is in the tools used. Instead of questionnaires, we construct samples of rich AIO data, and employ the most effective techniques for predictive analysis. By using AIO data and by predicting behavioural actions, such as political giving, we maintain the integrity of AIO profiling while utilizing data-driven techniques to detect trends that would otherwise go unnoticed by humans.

3.3.1 Ontology for AIO Profiling Text Analysis Using N-Grams

Twitter records consist of 140 character "microblog" statements, or "tweets" concerning a user's life experience. By treating a user profile's complete collection of tweets as a document, we can perform Natural Language Processing to identify features of the social data. By labelling each profile according to their political giving or affiliation, we can perform machine learning to identify the relationship between AIO and political giving data features. An advantage of machine learning over conventional human developed expert systems is that it allows machines to identify patterns or data features that would otherwise go undetected.

If we observe tweets, we realize that there are many ways that a tweet can be processed. Observe the following tweet:

M "It's raining but I don't believe that it is raining."

One approach to performing Twitter analysis on a word like this is to assess the semantics of the phrase. To a human, one might recognize that this tweet refers to rain and belief, and a reasonable observer might conclude that the Twitter user isn't making much sense. After all, if someone say's that "it's raining", they certainly believe that it is raining—in a sense the second half of the sentence contradicts the first. Using a computer, we might also be able to recognize this relationship by drawing on semantic reference resources such as WordNet [47]. By somehow linking this tweet to the meaning of the elementary terms, we could theoretically draw conclusions about this individual's personality and from that, an AIO profiling method.

However, this approach fails to capture the nuances of the relationship between terms and semantics. The meaning of phrases and terms are not atomic; they rely

on the linguistic context of the discourse. In this case, the meaning of the tweet is not related to eccentric utterances about rain at all. This phrase is often cited in the world of academic philosophy as "Moore's Paradox"; a linguistic tool to make a broader philosophical point about the normative relationship between truth and belief from the first person [33]. This case nicely demonstrates the challenges of semantic mining approaches using lexicons or sentiment analysis. This was not an eccentric tweet, but the utterance of some philosopher. Regardless of the quality of the library, semantic approaches to natural language processing are constantly challenged by the ambiguity of language.

Another approach could be to identify word unigrams, also known as "keywords", that might be characteristic of the desired profiles and to look at the numeric relationship between them and labels. Conover et al. [16] used a series of common political hashtags and user labels to identify political party affiliation. These hashtags were selected based on expert domain knowledge about politics, and were related to profiles manually labeled by the same domain experts.

Rather than using an expert approach, we leverage the value of our factual labelled data to create a word unigram technique that does not require manual labeling. Instead, we identify all tweets that occur more than 4 times in the sample dataset. The relative frequency of these hashtags are then used to build a probabilistic predictive model, which is in turn tested for effectiveness. This "bag of words" approach is commonly used to solve these sorts of classification problems. By creating a dataset which summarizes the word unigram mentions, we might assess the relationship between mentions and political giving, similar to Conover et al. [16].

A challenge to this approach is that it largely ignores the majority of data and only utilizes content tokens to draw larger conclusions about a user's personality. As we know, relevant factors in someone's personality, such as activities, interests or opinions, are not easily reducible to a collection of key terms contained in a small portion of relevant texts. Even by creating a substantial collection of relevant hashtags or terms, we would be largely unable to account for the majority of data that might be implicit in other aspects of a Twitter profile.

In response, we could evaluate relationships between character n-grams instead. Character n-grams are collections of characters common throughout an author's

works. Previous explorations with character n-grams have been successfully utilized in authorship attribution [43] and financial forecasting [9] and could be applied to detect unforseen relationships between authors who have political affiliations or are likely to make political donations. Importantly, character n-grams make use of a much larger body of text, and maintain the integrity implicit in common trends among user interests or personality. For example, collections of bigrams are useful at performing author attribution but also useful for detecting spam emails. If users share a common collection of terms or word use, even if psychological in nature, it might be captured implicitly in analysis with character n-grams.

3.3.2 Performance Evaluation

The success of psychographic mining tools can largely be attributed to their results. These tools lend predictive power to marketers who want to be able to understand their customers based on their motivations or behaviours. However, the rich psychographic mining tools such as VALS or LOV are also extremely useful for their ability to add explanatory power to the desires and motivations of consumers. Marketers use them to understand how their customers "tick".

A data driven psychographic mining tool should not only be able to make predictions, but also be able to offer robust insight into the motivations of a firm's customers. As such, good psychographic mining tools should be able to make accurate predictions, and therefore perform well at classification tasks. Perhaps the primary advantage of data driven psychographic mining over traditional mining is its capacity for performing experiments. Given our donations dataset, data driven psychographic profiling should therefore be able to perform well at predicting political affiliation, and whether a donor has made a gift in the past.

Profiling success is measured by predictive performance and insight robustness. We can evaluate the success of various techniques by their ability to predict better than others. Given that there are multiple n-gram techniques, we measure success of character n-grams by measuring the effectiveness versus other techniques, in addition to the performance compared to baseline findings from related literature. Cohen et al. [14] find that it is challenging to make predictions about political affiliation using data from casually affiliated donors with accuracy over 65%. Given that a primary

advantage of our mining from disparate sources is the insight from non-vocal sources, we can use this as a benchmark for success. We can then evaluate the success of the profiling technique by its success versus hashtag analysis or the word unigram "bag of words" approach.

Chapter 4

Experiment Design

This chapter explores the design of the experiment. Where the previous chapter identified methods for conducting an experiment, here the data and technical details of the experiment are explored in greater depth. The experiment involves building classifiers for two tasks: Predicting Affiliations and Predicting Donations. In addition, four techniques are explored in each task, resulting in eight experiments. The details of these experiments are provided.

4.1 Techniques and Hypotheses

Keeping with the vision for a two-dimensional profiling scheme, we have two classification tasks to validate. The first class concerns whether an individual is likely to give a political donation. The second concerns the affiliation of the political donors. Experiments thus have to be conducted to discover the most effective method for creating classifiers that consider these two dimensions.

We test three techniques that are derived from the literature. The first is predicting political donations and affiliation using hashtag word unigrams, along the lines of Conover et al. The second is a typical "bag of words approach" that takes popular word unigrams and uses them to make predictions. The third is to utilize character n-grams to predict affiliation, similar to an author attribution problem. Prior to the experiments, we can form a series of hypotheses about the relationship between the classifiers and the techniques:

H1 Hashtag Word unigrams can predict political affiliation with 65% accuracy, along the findings of Conover et al.

H2 Optimal classification tasks using hashtag word unigrams will perform more accurately than word unigrams.

H3 Machine learning will be less apt to build models on Affiliation than Donations.

H4 Character N-Gram (CNG) techniques will outperform hashtag techniques.

H1 is derived from the principle that matching the FEC donations data is at least as accurate as manually classifying Twitter users. We hypothesize H2 is derived from the research of Conover, which indicates that hashtags should perform better than "bag of words" more generally. H3 is derived from the fact that the third "unknown" class of the political donations will create noise that will challenge meaningful classification. H4 is based on the performance of CNG versus other common techniques, such as the "bag of words" approach.

To test these hypotheses, there are eight experiments to perform. The first two concern using hashtag word unigrams to predict political affiliations and propensity to give. These experiments can establish whether there is consistency between the Cohen findings and our disparate data. Two other experiments can be conducted to evaluate the usefulness of a "bag of words" approach on predicting either political affiliation or donation propensity. Using a predictive tool such as Weka, we can evaluate the results using machine learning techniques such as Support Vector Machines, establishing a baseline effectiveness with this dataset. The four final experiments can utilize the CNG technique to establish an effective psychographic profiling scheme. The experiments utilize the hashtag and aggregate data, so that results can be compared with the first experiments. In this chapter, I discuss the data preparation and experimental process for each of the six experiments. Following this, the conclusions and analysis described in greater detail in Chapter 5.

4.2 The Data

The data for the experiments were retrieved from three distinct sources: FEC Filings, Twitter Profile Extracts and Tweets. FEC filings consist of 210 447 transaction records supplied to us by an industrial partner who keeps cleaned records of FEC donations. As described in Chapter 3, the FEC Filings consist of transaction records of donation amounts, political parties given to and the name and addresses of the donors. Tweets, by contrast, consist of a series of short utterances about the user's experience. Referring back to the summary table, the first task was to prepare the experiment data for integration.

Using the name features of the FEC and Twitter Profiles, we are able to perform integration using MySQL. In addition to the FEC filings, an industry partner provided 119071 Twitter records containing user names and profile summaries from the US state of Missouri. Table 4.1 describes what was provided.

Data Feature	Type	Description
User_Id	DOUBLE	Unique Twitter user id
RealName	TEXT	Self-declared real name
Screen_name	TEXT	Twitter profile name
Buying_Stage	BIGINT(20)	Buying stage identified by partner
Followers	DOUBLE	Number of followers
Friends	DOUBLE	Total friends
Lifestyle	TEXT	List of lifestyle segments identified by partner
City	TEXT	User's self-declared city
State	TEXT	User's self-declared state
Country	TEXT	User's self-declared country

Table 4.1: Twitter Features

Using the Python nameparser library [27], the real names for each dataset were parsed into first and last names and added to MySQL. Using the first name, last name and city features, the FEC summaries were integrated with the Twitter features provided. This resulted in 480 matches. Each match was then manually evaluated for plausibility based on occupation criteria. If a Twitter profile was indicative of a nonmatch based on occupation, the profile was discounted. This resulted in a total of 219 users from whom we could be confident about the match. For cleaning ease, the data was then exported to csv where unusual characters could be removed using RPython. This results in a set of labelled Twitter profiles that can be compared using whatever attributes we might desire.

The next phase involves extracting and processing user tweets for analysis. This seems comparatively daunting, as though each profile might contain a comparatively limited number of FEC records, and each user might have up to hundreds of thousands of tweets. When we consider that each tweet consists of up to 140 characters, the data is comparatively quite large when we realize that there are potentially hundreds of thousands of profiles to analyze. This is one of the advantages of restricting analysis to a specific geographic region, such as Missouri.

The Twitter API is used to extract Tweets, which are in turn processed using a

program designed to test each of the techniques. Using the Tweepy library [32], we are able to construct the poli_tweet_dumper.py Python program that reads a list of Twitter users, call the Twitter API to collect Tweets for that user, and record the tweets in CSV format. Modified from code provided publicly by the tweepy community, the poli_tweet_dumper.py program was constructed to extract the precise tweets, while also considering the API query limits. Profile data was extracted from the 219 clean users. In addition, 219 profiles were randomly selected from the Missouri Twitter profiles to serve as a control for political donation detection. Each of these profiles were manually investigated to ensure that they neither belonged to the political sample, and were not spam accounts.

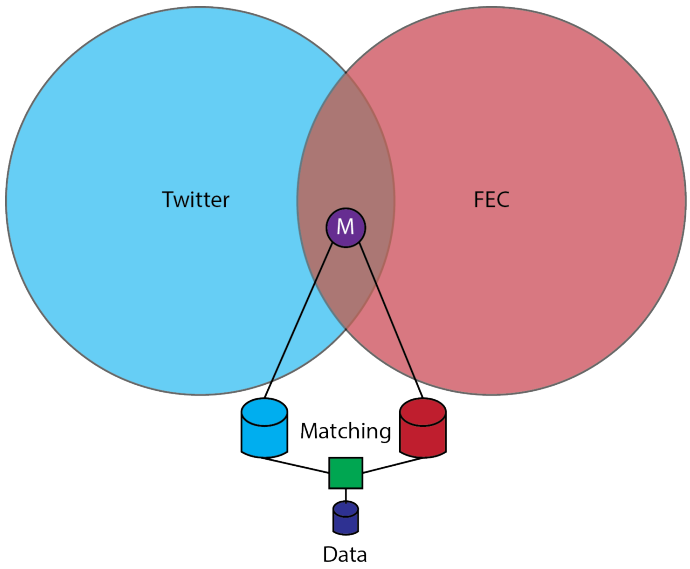

Figure 4.1: Political Affiliation with Hashtag Unigrams

The extracted tweets were saved in two folders: one for the political donors and one for the control group. The labels for each of the political donors and nondonors were saved in the same folder, and later appended to the N-Gram summaries and using R. All predictive experiments used these merged political files and Weka, save the character N-Gram experiments, which used the Ngrams.pl Perl script and CNG module designed by Dr. Vlado Keselj at Dalhousie University.

4.3 System for Extracting Hashtag and Word Unigrams

The first experiments involves testing the effectiveness of conventional techniques using the integrated dataset. Two techniques are derived from the literature and tested for predictive success. The first technique utilizes hashtag word unigrams extracted from a selection of most popular hashtags. The second technique utilizes word unigrams with no particular restriction. Tweets from user profiles were extracted and summarized using the hashtags.py script.

The hashtags.py script script aggregates the collective tweets from the sample, and using Python's Natural Language Processing Toolkit (NLT) it extracts a series of words that meet specific criteria. In the case of these experiments, all hashtags or words that occurred at least five times in the total sum of tweets were selected for evaluation. The hashtags.py program then evaluates each user's tweets individually, and aggregates the number of occurrences of each word unigram. The program writes each aggregation as a row in a CSV filesheet, producing a sheet with a number of data features and the number of times each feature occurred.

Following the creation of this CSV filesheet, the findings are aggregated with the labels generated from the FEC filings, matching each user according to their Twitter screen name. Using R, the matching is performed, cleaned and rewritten in csv form. This labelled CSV data is then evaluated using a machine learning tool such as Weka.

4.3.1 Predicting Political Affiliations

The Affiliations data consists of the tweet records from 219 individuals for whom we have FEC matches. The Tweet records are aggregated and labelled according to the technique in question. Each dataset is then evaluated using a series of predictive techniques. All predictive techniques utilized in this experiment were classification tasks, which seek to identify a user as making donations to "DEM" (Democrat), "REP", (Republican) or "UNK" (Unknown). The "DEM" class contained 68 profiles, while there were 64 "REP" entries and 88 "Unknown" profiles. The table below describes the number of data features extracted from the tweets collected from the affiliation dataset.

The predictions were conducted using three predictive techniques known for their

Technique	Number of Features
Hashtag Word Unigrams	1991
Word Unigrams	24671

Table 4.2: Features of Affiliation Experiments

ability to process sparse data: Support Vector Machines (SVM), Naïve Bayes and the C 4.5 Entropy Tree. Each utilizes the data features to make predictions, so the number and quality of features determines the predictive power. In order to determine the optimal number of features, the features are trimmed and tested against each technique. In order to ensure adequate accuracy of scale, features are trimmed on exponential basis, on a factor of two. By doing this, and repeating the trimming process for each technique, we ensure that the potential of each technique is better realized. It is important to note that the number of attributes for the word n-gram database is substantially larger than those from the hashtags. To help ensure methodological consistency, the trimming was performed incrementally.

4.3.2 Predicting Donations

Much like the experiments with Affiliations, Donation experiments are conducted using collections of word hashtag unigrams. The classification task is different however. Whereas Affiliations concerned the affinity to a particular political party or cause, Donations classification tasks concern whether an individual has the propensity to make a political donation. Our intuitions suggest that politically oriented Twitter users could exhibit traits, such as political activity, strong opinions or displays of wealth, that can be used to predict whether someone is apt to make a donation. Machine learning can therefore be performed to measure this vector in the psychographic profiling scheme.

To test whether an individual is likely to give, we created a binary Donations class, based on whether a user record was contained in the FEC data. Tweets from the 219 FEC donors were compared with 219 profiles randomly selected from Missouri. Like with the Affiliation data, two classes were created: “True” indicating a match in FEC and “False” indicating no match. The counts from hashtag and word unigrams were compared again using Naïve Bayes, Support Vector Machines and C4.5 Entropy Trees. The number of features from the Donations experiments are described below.

Technique	Number of Features
Hashtag Word Unigrams	3368
Word Unigrams	36647

Table 4.3: Features of Donations Experiments

Like with the Affiliation experiments, the Donations experiments involved attribute trimming on the "N/2" scale. Given that the sample size was larger, the number of word and hashtag unigrams were substantially larger. This resulted in a larger number of experimental rounds.

4.4 System for Performing Character N-Gram Analysis

In addition to word unigrams, we examined the feasibility of using Character N-Grams (CNG) for the classification tasks. Unlike word unigrams, CNG utilizes the tweet's characters to produce a series of distance models on a variety of character n-gram variations. Character unigrams, bi-grams through to 10-grams were evaluated through the course of these experiments. In addition, the script considers multiple attribute models for prediction, ranging from the first 20 n-grams to the first 10 000. Both Affiliation and Donor experiments were conducted accordingly. Unlike the previous experiments, the CNG experiments were conducted completely in Perl using the CNG module designed by Dr. Vlado Keselj and the Dalhousie Natural Language Processing (DNLP) lab.

Data cleaning and preparation were completed for processing using Perl and bash scripts. As with the word unigrams, the sample data consisted in a series of tweets. For the Affiliation experiments, the tweets from the 219 sample users were collected and divided into test and train folders. 146 random profiles were included in the test folder, while 73 files were included for training. Using the Perl scripts, the tweet extracts were cleaned for timestamps and metadata, leaving only the tweets behind. Once executed, the CNG script builds a series of n-gram training models on the training data and measures success of that model against the classes. In the affiliation dataset, as before, the classes consist of a relatively even distribution of "UNK", "DEM", "REP" classes.

We conducted two experiments on the Affiliation dataset. The first utilized word

unigrams collected from the user tweets. The collective tweets were aggregated and assessed by class according to a series of n-gram models generated from collections of tweets. Following this, the datasets were cleaned to only include hashtag and user mentions. A second experiment was conducted on this hashtags dataset, to determine a relationship between hashtag mentions and political affiliation. These two steps were again repeated with the 438 records from the Donations dataset. Finally, experimental results were recorded according to the n-grams and number of grams used to build the most effective model.

Chapter 5

Results and Discussion

This chapter explores the results of the experiment, and discusses the findings in the context of the original research problem. Using Naïve Bayes and CNG, we were able to build predictors that accomplished both of the two classification tasks with greater than 15% accuracy over the respective majority classifiers. This implies that we are able to make meaningful predictions about FEC filings using Twitter. Using classification techniques, we can therefore build a psychographic classification system based on the desired features. This chapter begins by exploring the results from the experiments. It then revisits the original experimental hypotheses described in Chapter 4, followed by a description of the technique's unique marketing utility. It concludes with an exploration of the method's limitations.

5.1 Evaluation of Word and Hashtag Unigrams

The original task of evaluating predictive success of word and hashtag unigrams on Political Affiliation found that word n-grams were largely ineffective for building predictive models for Affiliation. On word unigrams particularly, no model was able to perform better than the majority classifier of 38%. Hashtag unigrams, which performed well with the Conover data, were not much more helpful for building a good model. The best classifier using aggregated word unigrams from Twitter achieved 38% accuracy, while the entropy trees managed to achieve 42% accuracy. This suggests that the process of using hashtags instead of the aggregate words might prove beneficial from the reduction of noise, but the 4% improvement is still significantly below the accuracy suggested by the literature. The graphs below demonstrate the results of the political affiliation experiments, and compares the best results of the hashtag and word unigram experiments.

These results were disappointing. We were largely unable to repeat the findings of Conover or Cohen. Initial investigation of the donor word unigrams proved equally

	Features					
	62	124	249	498	995	1991
C4.5	42	40.6	33.3	37	41.6	37
NaiveBayes	32.9	37	37.4	33.3	34.2	31.5
SVM	36.1	37.4	37.9	39.7	39.7	37.9

Figure 5.1: Political Affiliation with Hashtag Unigrams

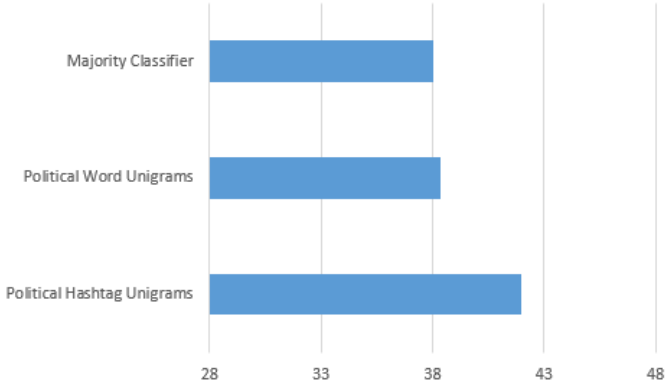

Figure 5.2: Comparison of Strongest Political Affiliation Results

disappointing. The Donations predictions using word unigrams were unable to produce a classifier that exceeded 52% accuracy, or two percentage points beyond the majority class. However, hashtag unigrams performed significantly better, generating a model using Naïve Bayes with 3368 attributes and over 66% accuracy. This is over 16% above the majority classifier, and is in line with the expectations suggested by Cohen et al, which states that mining political affiliation over Twitter might regularly exhibit no more than 65% accuracy. The graphs below demonstrate the success of the various models. Note that in the case of donor hashtags, all three techniques performed strongly.

	Features						
	52	105	210	421	842	1684	3368
C4.5	55.5	59.1	59.6	64.2	63	63	63
NaiveBayes	56.4	56.8	55.3	60	62.8	63.5	66.2
SVM	53.2	56.4	56.8	60	61.2	64.4	64.2

Figure 5.3: Donor Propensity with Word Unigrams

The success of the hashtag word unigrams over regular word unigrams suggests that there is a significant amount of noise within Twitter data, and that this noise can largely be attributed to the fact that political conversation makes only a small portion

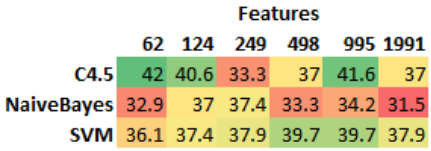

Figure 5.4: Donor Propensity with Hashtag Unigrams

Task	No Classes	Attributes	Technique	Accuracy
Political Hashtags	3	62	C4.5	0.42
Political Words	3	3082	C4.5	0.38
Donor Hashtags	2	3368	Naïve Bayes	0.66
Donor Words	2	263	SVM	0.52

Table 5.1: A Comparison of Optimal Word Unigram Models

of politically active Twitter users' conversation. In fact, simply observing a sample of the most popular hashtags and word unigram sheds insight into the problem of noise. Table 5.1 summarizes the best results from each of the experiments conducted on word unigrams.

5.2 Evaluation of Character N-Gram Models

Following the success of the Naïve Bayes on hashtag word unigrams, we determined the benefits of contrasting the CNG experiments on collections of words with collections of hashtags and mentions. On the Political Affiliation dataset, CNG technique was performed, and produced promising results. A model was able to accurately predict affiliation with 51% accuracy using character unigrams. Compared to previous results from the aggregated word dataset, which could not exceed the majority classifier of 38%, this is was extremely promising.

Perhaps even more encouraging were results from the cleaned hashtag and user mentions dataset. Using this set, a predictive model was constructed using 500 character bi-grams. Considering that the majority class "UNK" contributes to implicit noise in the classification system, the fact that a model was able to accurately classify such a large number of cases suggests that there is considerable grounds for performing more detailed assessment of political affiliation using CNG.

When applied to the Donor classification, though CNG was able to produce models with some predictive power, it was unable to exceed the 66% demonstrated using

N \ L	20	50	100	200	500	1000	1500	2000	3000	4000	5000	6000	7000	8000	9000	10K
1	0.37	0.37	0.38	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
2	0.23	0.27	0.3	0.38	0.45	0.45	0.42	0.42	0.4	0.38	0.44	0.42	0.44	0.44	0.44	0.44
3	0.33	0.3	0.38	0.4	0.36	0.4	0.37	0.4	0.44	0.45	0.41	0.44	0.44	0.45	0.49	0.47
4	0.23	0.33	0.41	0.29	0.42	0.38	0.36	0.37	0.41	0.42	0.36	0.41	0.37	0.38	0.37	0.4
5	0.3	0.33	0.23	0.33	0.38	0.32	0.3	0.34	0.3	0.36	0.37	0.32	0.34	0.33	0.34	0.34
6	0.32	0.27	0.32	0.36	0.3	0.32	0.4	0.37	0.41	0.37	0.38	0.36	0.33	0.36	0.4	0.34
7	0.29	0.34	0.34	0.38	0.27	0.34	0.37	0.34	0.37	0.4	0.42	0.41	0.4	0.38	0.37	0.36
8	0.23	0.29	0.3	0.3	0.36	0.33	0.33	0.29	0.4	0.42	0.4	0.41	0.36	0.38	0.36	0.36
9	0.22	0.27	0.26	0.27	0.3	0.27	0.32	0.32	0.37	0.37	0.33	0.32	0.3	0.29	0.32	0.32
10	0.26	0.22	0.3	0.3	0.29	0.27	0.34	0.33	0.3	0.29	0.27	0.32	0.29	0.33	0.32	0.34

Figure 5.5: Results for CNG Political Affiliation on Words

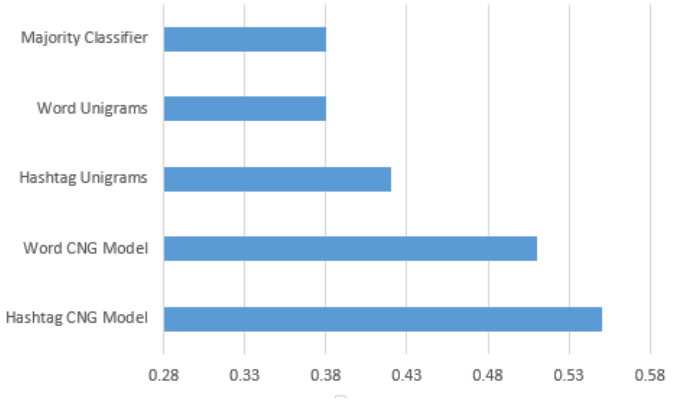

Figure 5.6: A Comparison of Strongest Political Affiliation Findings

N \ L	20	50	100	200	500	1000	1500	2000	3000	4000	5000	6000	7000	8000	9000	10K
1	0.34	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.4
2	0.27	0.34	0.34	0.38	0.55	0.48	0.49	0.44	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.4
3	0.34	0.42	0.47	0.44	0.4	0.49	0.45	0.44	0.48	0.49	0.48	0.47	0.47	0.47	0.47	0.5
4	0.3	0.34	0.32	0.41	0.44	0.47	0.51	0.51	0.49	0.48	0.48	0.47	0.49	0.51	0.51	0.5
5	0.36	0.42	0.34	0.41	0.47	0.53	0.47	0.47	0.47	0.52	0.49	0.51	0.52	0.51	0.52	0.5
6	0.33	0.37	0.4	0.44	0.51	0.55	0.44	0.47	0.44	0.47	0.48	0.52	0.49	0.51	0.53	0.5
7	0.36	0.36	0.42	0.44	0.49	0.47	0.52	0.48	0.44	0.47	0.49	0.51	0.53	0.52	0.51	0.5
8	0.36	0.37	0.38	0.49	0.48	0.45	0.49	0.51	0.48	0.48	0.48	0.51	0.51	0.49	0.49	0.5
9	0.34	0.34	0.44	0.47	0.51	0.47	0.49	0.51	0.48	0.51	0.51	0.49	0.48	0.52	0.49	0.5
10	0.34	0.33	0.4	0.49	0.49	0.49	0.53	0.52	0.52	0.52	0.51	0.51	0.49	0.49	0.49	0.5

Figure 5.7: Results for CNG Political Affiliation on Hashtags

hashtag word unigrams. Though the results from these experiments were promising and were able to yield classifiers that modeled 60-61% accuracy, they did not seem to capture some of the qualities implicit in the word unigram hashtags.

N \ L	20	50	100	200	500	1000	1500	2000	3000	4000	5000	6000	7000	8000	9000	10K
1	0.5	0.58	0.54	0.54	0.54	0.54	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55
2	0.47	0.49	0.49	0.58	0.54	0.53	0.52	0.51	0.5	0.53	0.53	0.53	0.53	0.53	0.53	0.53
3	0.51	0.52	0.6	0.53	0.52	0.51	0.49	0.47	0.52	0.47	0.48	0.49	0.49	0.49	0.48	0.48
4	0.46	0.52	0.53	0.53	0.49	0.51	0.49	0.51	0.51	0.47	0.49	0.53	0.51	0.46	0.52	0.52
5	0.55	0.55	0.58	0.59	0.49	0.5	0.49	0.47	0.47	0.47	0.5	0.53	0.49	0.47	0.48	0.49
6	0.53	0.53	0.56	0.57	0.45	0.48	0.55	0.51	0.5	0.51	0.51	0.49	0.53	0.55	0.51	0.53
7	0.53	0.53	0.55	0.51	0.47	0.49	0.53	0.54	0.49	0.48	0.47	0.51	0.55	0.52	0.49	0.51
8	0.53	0.53	0.52	0.51	0.47	0.52	0.49	0.49	0.52	0.54	0.52	0.53	0.53	0.51	0.51	0.53
9	0.53	0.53	0.52	0.53	0.47	0.51	0.48	0.51	0.57	0.54	0.56	0.55	0.55	0.55	0.54	0.51
10	0.53	0.53	0.52	0.53	0.51	0.49	0.51	0.51	0.51	0.53	0.51	0.51	0.52	0.54	0.52	0.52

Figure 5.8: Results for CNG Donations Propensity on Words

N \ L	20	50	100	200	500	1000	1500	2000	3000	4000	5000	6000	7000	8000	9000	10K
1	0.5	0.58	0.53	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54
2	0.5	0.57	0.49	0.57	0.53	0.49	0.5	0.53	0.54	0.53	0.55	0.53	0.53	0.53	0.53	0.53
3	0.53	0.47	0.56	0.53	0.53	0.52	0.49	0.49	0.48	0.51	0.49	0.47	0.5	0.5	0.5	0.45
4	0.51	0.58	0.54	0.49	0.51	0.51	0.53	0.51	0.51	0.53	0.51	0.55	0.56	0.55	0.53	0.53
5	0.57	0.42	0.51	0.57	0.56	0.54	0.5	0.52	0.53	0.55	0.53	0.55	0.56	0.54	0.54	0.55
6	0.52	0.49	0.53	0.53	0.53	0.53	0.55	0.53	0.56	0.57	0.55	0.6	0.6	0.53	0.53	0.55
7	0.53	0.52	0.5	0.51	0.51	0.5	0.53	0.59	0.55	0.58	0.58	0.56	0.54	0.54	0.55	0.54
8	0.58	0.53	0.53	0.48	0.51	0.55	0.58	0.59	0.56	0.56	0.55	0.53	0.53	0.51	0.53	0.51
9	0.58	0.53	0.52	0.44	0.49	0.61	0.6	0.58	0.55	0.57	0.53	0.55	0.54	0.53	0.53	0.54
10	0.53	0.5	0.49	0.45	0.51	0.57	0.59	0.52	0.55	0.53	0.53	0.54	0.51	0.55	0.56	0.55

Figure 5.9: Results for CNG Donations Propensity on Hashtags only

5.3 Hypotheses Revisited

If we revisit the hypotheses outlined at the outset of the experiment, we realize that the evidence lends significant support some hypotheses, while challenging others. As we recall, the hypotheses were as follows:

H1 Hashtag Word unigrams can predict political affiliation with 65% accuracy, along the findings of Cohen et al.

Classification Task	No Classes	N	Length	Accuracy
Affiliation Word Unigrams	3	1	1000	0.51
Affiliation Hashtags	3	2	500	0.55
Donor Word Unigrams	2	3	100	0.60
Donor Hashtags	2	9	1000	0.61

Table 5.2: A Comparison of Optimal N-Gram Models

H2 Optimal classification tasks using hashtag word unigrams will perform more accurately than word unigrams.

H3 Machine learning will be less apt to build models on Affiliation than Donations.

H4 Character N-Gram (CNG) techniques will outperform hashtag techniques.

The results from our investigation of hashtag unigrams challenged the supposition of H1. Using hashtag unigrams and the FEC data, we were only able to classify users with 42% accuracy. Understanding that our sample data is dissimilar from the data tested by Cohen et al (we have a third "unknown" class, while Cohen et al had only two), the fact that our labels came from a concrete source of ground truth casts some skepticism about using domain experts to classify political data. The fact that CNG was able to generate that was significantly better than word unigrams further suggests that the process of labelling using FEC data was not flawed, and that the matching process has merit. Though this requires further investigation, this supports Cohen's skepticism about the method of using domain experts to generate political classifiers.

H2 was largely supported by the findings. In both word unigrams and CNG, models generated from hashtag data consistently performed better than data generated from word aggregates. However the significant improvements of using hashtag word unigrams on Donor experiments were not replicated in the Political Affiliation experiments. We can speculate that this was because of the noise generated from the "unknown" class in the political affiliation experiments.

H3 was originally conceived because of two reasons: that the political affiliations data has three classes and because the affiliations dataset was half of the size of the Donations data. However, the fact that CNG was able to generate a model that was 17% above the majority classifier on this dataset challenges both of these assumptions. It could be that the social traits that differentiate political affiliations are much stronger than those that determine the sorts of people who donate to political parties. Further investigation into this factor is required.

Finally, H4 is weakly supported by the findings. The CNG classifier performed far better than any of the hashtag unigrams when determining political affiliation, however, CNG did not perform as well as the Naïve Bayes model when determining donor status. The fact that CNG consistently performed well however, often far

exceeding the word unigram models suggests that the application of CNG to political donation problems. These findings also suggest the utility of applying CNG to data driven AIO mining more generally.

5.4 Marketing Utility of AIO Mining of Disparate Datasets

The model constructed using CNG and Naïve Bayes provides a foundation of a broader method for robust data-driven psychographic research. Using the two vectors, we are able to determine with some level of accuracy whether a given Twitter profile is likely to have a political affiliation, and whether they have a propensity to give to political parties. Table 5.3 describes the results of the best classification models for each task, along with the majority classifiers.

AIO Vector	Experiment	Technique	Accuracy	Majority Classifier
Political Affiliation	Hashtags	CNG	0.55	0.38
Donation Propensity	Hashtags	Naïve Bayes	0.66	0.50

Table 5.3: Insights from Optimal Techniques for Identifying AIO Vectors

These classifiers represent AIO vectors in a wider psychographic classification scheme. What do these vectors look like? The vectors that resulted from the experiments not only incorporate features of text, but also robust data about users' activities, interests and opinions. Each vector incorporates features from Twitter data and incorporates the results of an action label. This is not fundamentally different from the other psychographic methods. Consider a sample of the model generated from the Donations classifier in Table 5.4.

Hashtag	#god	#stemcells	#brain	#cardswarmup
True (std. dev)	4.1517	3.5826	69.2433	2.3611
False (std. dev)	1.4524	2.4475	4.371	13.0609

Table 5.4: Sample of Naïve Bayes Model Features with Strong Determinants

The Naïve Bayes model makes predictions based on the probability of a certain class belonging to a specific label. It establishes the probability based on the differences in the labels between the sample data and the mean. In this case, the model differentiated the classes using hashtags such as "#god", "#stemcells" and "#brain",

associating them with people who have made political contributions, which hashtags such as "#cardswarmup" were identified with individuals who did not make donations. We notice that the occurrence of certain data features, such as "#brain" are more common in one class versus the other. We can tell this from the standard deviation values associated with each feature. High standard deviations associated with one of the classes suggest that data with these features are likely to be classified with a particular class when applying this model. The result makes sense, as the first three hashtags mentioned are associated with topics that are politically contentious in Missouri, such as religion and life issues. The "#cardswarmup" hashtag by contrast is associated with the charitable status of the St. Louis Cardinals baseball team, and has little to do with political activity. This makes sense given that the sports team likely appeals to a more general population that does not necessarily give to political parties.

However, the model also provides unexpected insight into the AIO features of political donors in particular. Consider the table below. These results are surprising. The "#moleg" hashtag is the designated hashtag of the Missouri Legislature, the governing body for the state of Missouri. "#voteblue" is a hashtag encouraging support for the Democratic party, "#earthhour" is an environmental movement, and "#progress" is common hashtag used in progressive discourse. All of these hashtags are highly political, but are not strong determinants of political donations. This is very interesting, as conventional wisdom would put weight on political terms over terms such as "#amazing", a hashtag which is strongly associated with political donors. These findings offer significant insight into the AIO profiles of the individuals who give to political parties. It seems that political activism is not necessarily a strong determinant of donations, as if they were more commonly associated with a particular class in the training data, the differences in standard deviation would be more profound.

Hashtag	#moleg	#voteblue	#earthhour	#progress
True (std. dev)	0.4167	0.3333	0.3248	0.4045
False (std. dev)	0.4167	0.4258	0.1667	0.3333

Table 5.5: Sample of Naïve Bayes Model Features with Weak Determinants

Concerning character n-grams, the models offer very interesting insight that would

not otherwise be available to us. The most successful n-gram model consisted of bi-grams, or combinations of two characters that are used to determine distance for the classes. We can extract the bi-grams from the successful models and compare how the classification logic works. Table 5.6 compares the ten most common bi-grams from the REP, DEM and UNK classes.

DEM	Distance	REP	Distance	UNK	Distance
\n @	0.03765	\n @	0.03792	\n @	0.03588
_ @	0.02170	_ #	0.01867	_ @	0.02182
_ #	0.01913	_ @	0.01646	_ #	0.01869
e r	0.01287	\n #	0.01212	e r	0.01376
a r	0.01062	e r	0.01175	a n	0.01119
o n	0.00978	a r	0.01064	o n	0.01061
a n	0.00904	a n	0.01056	a r	0.01007
\n #	0.00844	s \n	0.01045	\n #	0.00899
i n	0.00727	i n	0.01016	i n	0.00856
s t	0.00700	o n	0.00868	l e	0.00741

Table 5.6: Ten Popular Bi-Grams from Each Class

Though the DEM and REP n-grams share many things in common, there are significant differences in the distance between even the ten most popular n-grams. DEM, for instance, are far more likely to use the "_ @" n-gram, which is indicative of a mention of another Twitter user. We might infer from this that DEM users are more sociable and can do further investigation on the social differences between DEM and REP. Interestingly, REP users from our set are more likely to use the the "\n #" bigrams, which is indicative of using a hashtag on a new line. These subtle differences in behaviour would have gone unnoticed using word n-gram models, and give grounds for forming new hypotheses about behavioural patterns among DEM and REP users on Twitter. These can be further investigated to extract marketing insight about the users.

Perhaps the strongest feature of our data-driven method is its resistance to the "confirmation bias" created by domain experts, as alluded to in the Cohen paper. Where domain expert generated methods for political and psychographic profiling focus on a collection of Twitter profiles and features that are subject to bias, this method is able to resist the bias by extracting labels from a disparate source. The result is in a model that is at times extremely unintuitive, but able to generate

robust insight into the activities and interests of users outside of the specific domain in question. Using this model in political giving as a foundation, the profiling method could be expanded into other domains, ultimately concluding with a comprehensive profiling tool of a similar nature to VALS.

By using these two models as classifiers, we are also able to generate insight from Twitter about the attitudes of collections of individuals with these psychographic profiles. By extracting a series of Tweets from a sample of users, we can use these profiling tools to label the users and observe their reactions to certain events or products. For instance, by collecting a sample of 50 000 Twitter profiles and using these profiling classifiers to label the users, a marketing research agency might be able to measure the impact of a political campaign on the attitudes of individuals who are likely to make donations or share political affiliation. The advantage of using these classes as labels, rather than say popular hashtags or a life stream of tweets is that the labels implicitly contain AIO features that have been empirically verified, in part, by a ground truth external to Twitter.

Applied to domains outside of political giving, we might follow the same method to generate psychographic profiles for consumer products. Using retail data, we might be able to perform record linkage on a small subsection of users and label them accordingly. Using these labels, classifiers can be constructed and "multivector" profiles might be produced. The result could be in a much more comprehensive data driven solution that leverages Twitter data to produce comprehensive consumer insights.

5.5 Limitations of Methodology

5.5.1 Record Linkage

As stated, the fact that the record labels were extracted from disparate data helps ground the AIO models with a concrete gold standard. However, one of the greatest challenges to this method is that the record linkage, which connects the data from the FEC to the Twitter data, is not necessarily accurate. Matching was performed based on Geographic, Name and Occupation features, but even then, significant manual cleaning was required to ensure a degree of quality in the matching. A problem with record matching from disparate databases is that there is no way to truly ensure that

an accurate match has taken place. In addition, the fact that a large degree of manual matching was required limited the application of this method to a larger scale.

However, for firms or researchers with access to a good matching protocol, this might be less of an issue. Political parties, for instance, might request that donors might self-declare their Twitter information upon making a donation. Rewards programs might request members to link their profiles with their membership cards in exchange for a point bonus, yielding a reliable gold standard. In a research setting, we might conduct a comprehensive user study in which individuals are asked to link their profiles to the results derived in a laboratory. Regardless of the desired experiment, these sorts of matching techniques could be conducted to help ensure the integrity of the disparate data. With a reliable matching mechanism, the results from future experiments might greatly exceed the results explained in this paper.

5.5.2 Scope and Applicability

The method explained in this paper was limited in scope, specifically to political giving. Trends and findings in political giving might not translate into the wider world of psychographic profiling. The original ambition of this project was to outline a comprehensive profiling system that leverages disparate sources to generate psychographic insights. However, without sufficient time and labour resources, a comprehensive system could not be constructed. In addition, without comprehensive consumer data for labelling, a robust AIO profiling system could not be demonstrated.

The limitations of the political profiling could actually prove to be an advantage. Publicly discussing the details of customer profiles generated from existent marketing data could jeopardize customer confidentiality. In addition, profiling techniques generated from valuable customer data are in turn strategically valuable to firms, who might be reluctant to release proprietary data to the public. Much like Khale's LOV, the results of a psychographic profiling scheme based on publicly available political data could serve as a valuable learning tool for future researchers, without revealing the details of a potentially commercially valuable system. Future research could be undertaken as a commercial project, similar in spirit to VALS.

5.5.3 Political Affiliation Classes

A significant barrier to evaluating the effectiveness of the political affiliation classifier is that there are three classes ("REP", "DEM", "UNK") in comparison to the two classes specified in the rest of the literature. In addition, the third "UNK" class adds considerable noise to the result, creating a system that successfully classifies the "UNK" class as well as the other two. It is difficult to evaluate whether the affiliation classifier performs better than those described in the literature.

However, this problem is not unique to our case. Cohen et al. [14] describe the difficulties of recreating results from other datasets, citing a the inability to replicate domain expertise necessary to create the data. In our case, the fact that we were able to generate a classifier that was 16% above the majority class using three classes suggests that this could be a significant improvement over the state of the literature described in Cohen, which suggests it is maximally viable to build a classifier that is 15% above majority. In addition, our system is improvable by using a larger sample set, removing the "UNK" class, or by building better matching techniques.

5.5.4 Limitations of Scale and Sample

This experiment utilized data from 219 Twitter users who could be matched against FEC records from the state of Missouri. The total sample of Twitter users from Missouri was 119 071, thus only 0.18% of Twitter users could be matched against FEC filings. Considering that occupation data was among the criteria used for matching, the matching process might be heavily biased toward particularly vocal individuals who make their occupation information public. Considering that these might be individuals such as politicians, lawyers, businesspeople, and doctors, the classification model described in this paper might be heavily biased toward identifying certain vocal professionals.

This problem is not particular to this profiling model. According to the Pew Research Centre, in 2013 only 16% of American adults use Twitter [11] and of these, 52% of them actively use Twitter as a source for news. Of those active individuals, 45% are between the ages of 18 and 29, and 48% of them report incomes in excess of $75 000 and 40% report at least a bachelor's education. Active Twitter users represent a young, educated and wealthy demographic. Any marketing informatics

tool created using Twitter thus represents a small percentage of the wider population and should be recognized as part of a wider profiling scheme.

However, not all Twitter users fit this mold. By using a stronger matching technique and by taking a sufficiently large scale with attention to underrepresented demographic groups, this profiling technique can be applied to specific user sets to build profiles that identify Twitter users that are not otherwise easily identified using other models. Further research in this subject should incorporate details about the implicit bias of Twitter and include samples that help overcome these biases.

5.5.5 Privacy Implications

The system described in this paper involves record linkage from disparate data sources, and could be used to gain detailed insight about individuals. Data used in these experiments were extracted from publicly available sources of which individuals have consented to publication, which meets the expectations of research ethics in Canada. However, this technique could be used to extrapolate private information from non-public sets, or could be used to publish information on specific individuals without users' explicit consent. This might not only prove morally dubious but research conducted might violate the tri-council policy on research involving record linkage.

Proper ethical use of this system might be for firms to utilize data only for which they have explicit consent from users to use. In this respect, Twitter is a great tool for performing AIO mining, as users have given explicit consent to publicly provide their data. However, other commercial datasets acquired by web scrapers or web tracking do not necessarily acquire users' consent. Applying this or a similar technique to dubious datasets should be resisted.

Chapter 6

Conclusion and Future Work

We began this project with the goal of creating a data-driven solution to conducting activities, interests and opinion (AIO) research of a quality similar to that of VALS and LOV. The process took us from observing psychographic mining techniques to utilizing Natural Language Processing to generating detailed insights into the motivations of particular groups of consumers. We conclude the research not only by proposing a system that performs robust psychographic mining, but also grounds the psychographic profiles using a consistent empirical process based on predictive accuracy.

The use of matching using disparate datasets for generating labels has so far not been utilized by other researchers in the field of political affiliation or donations. Though we cannot draw conclusions for other domains, the criticism of Cohen and the failure of hashtag unigrams to produce comparable results to CNG suggests that there is considerable potential for the application of the CNG technique to affiliation problems. An advantage of using CNG and probabilistic models is that they also generate considerable marketing insight in themselves. Other advanced methods, such as neural networks, might not have the same robustness when applied to psychographic research.

The profiling system is also highly scalable. The system used in these examples used two vectors to create profiles: political affinity and political donation propensity. Additional vectors might be created in the same vein. Using consumer data, we might create profiles about bicycle enthusiasts, social media junkies, wine lovers or environmental consumers. Using other disparate classes such as transit records, governments might create useful tools for extracting public opinion of specific users who exhibit tendencies characteristic of transit riders, informing better policy decisions. The possibilities can be extended to create analytical tools of various specificities, broad or specific. The only limits are the set of quality classifiers, effective matching

processes and tangible problems.

It is our vision that the findings of this project might spark additional research in the domains of political prediction, digital democracy and marketing informatics broadly. In addition, the promising results of the novel method for dataset labelling and the application of CNG to political affiliation could offer a new area of application. The remaining sections of this chapter explore the potential directions described.

6.1 Dataset Extension for Political Affiliation N-Gram Analysis

One way to improve the results of the political affiliations vector is to expand the FEC dataset beyond Missouri, and focus on individuals whose donations match only the "REP" or "DEM" classes. If 219 quality matches could be created using data from Missouri, which represents roughly 2% of the US population, we should be able to extract roughly 20 000 quality matches from the sample of the entire United States. This sample could yield up to 12 400 quality profiles which could be classified according to their giving behaviours.

A larger sample set could not only offer more robust classification results, but allow us to gain deeper insight into the differences between the two political classes. Furthermore, by restricting the experiments to only two classes, results could be compared with the Conover and Cohen data to extend the field of political donations. By comparing the FEC filing results using hashtag unigrams, we could establish a confidence baseline that could either ground or criticize the practice of using domain experts for classification. If the hashtag unigrams do not reflect the results of the literature, this could cast doubt behind this process.

In addition, if CNG were to offer as dramatic an improvement over the hashtag unigrams as discovered in this project, it could offer a new standard for performing affiliation mining on Twitter. CNG seems to be uniquely well equipped to process noisy or non-obvious data, and establishing performance based on some degree of non-subjective truth could open a new frontier of online demographic mining.

6.2 Applications to Digital Democracy

Democracy was originally envisioned as a system for the empowerment of individuals in liberal societies. By casting ballots to elect representatives, democratic societies might call themselves just, insofar as they grant a system for giving voice groups that are not necessarily rich or powerful. In modern democracies, politicians use political polls to build policies that are in agreement with the interests of voters. By polling voters using media such as telephones, polling companies therefore play a critical role in the discernment of politicians and policy makers.

However, in recent years, political polls have come under criticism [78]. Polling companies, which rely on landline phones, have been challenged to create reliable results. As fewer young people own landline phones, polls have increasingly been skewed toward older individuals. A robust psychographic profiling system using Twitter could offer a new technique for conducting online polls, and one that can be created to match a range of specificity. Polling companies could utilize quality samples of social media users as a gold standard, and utilize their past behaviour as a sample for a classification system. Using this gold standard, they might be able to classify using the techniques outlined in this paper. With a more robust classifier, political opinions could be reliably mined for political impact.

6.3 Alternative Sources for Gold Standards

Finally, the larger project of a robust data-driven psychographic profiling system of a quality similar to VALS could be achieved using this method and a large amount of consumer data. As the commercial applications of psychographic mining are obvious, such a project would be best suited to enterprise, rather than the academy. VALS, which was originally conceived in the public research space of SRI International, was later refined in the private sector. In a similar vein, this type of research could be refined and applied to a larger project.

Such a company might be conducted in a matter that achieves a high degree of public good, rather than the privacy dystopia envisioned in science fiction. Using publicly available or ethically acquired marketing data to gain consumer insight could help to create a world where products are created that contribute the maximal value

to the consumers in question. In a world of better value products, there is less waste and better value chain management, benefiting society as a whole.

Bibliography

[1] G. Adomavicius and A. Tuzhilin. Using data mining methods to build customer profiles. *IEEE Computer*, February 2001.

[2] Leman Akoglu. Quantifying political polarity based on bipartite opinion networks. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[3] Environics Analytics. PRIZM C2. http://www.environicsanalytics.ca/data/segmentation/prizmc2, 2015. Accessed 12 March 2015.

[4] J. Aquino. Transforming social media data into predictive analytics. *CRM Magazine*, 16(11):38–43, 2012.

[5] Lisa M Austin. Reviewing PIPEDA: Control, privacy and the limits of fair information practices. *Canadian Business Law Journal*, 44:21, 2006.

[6] C. Axelrad. R.I.P. donor pyramid? http://www.fundraisingsuccessmag.com/article/rip-donor-pyramid/1, May 2014. Accessed 21 April 2015.

[7] Robert C Bartlett, Susan D Collins, et al. *Aristotle's Nicomachean ethics*. University of Chicago Press, 2011.

[8] T. A. Bennett and C. Bayrak. Bridging the data integration gap: From theory to implementation. *ACM SIGSOFT*, 36(3):36, 2008.

[9] Matthew Butler and Vlado Kešelj. Financial forecasting using character n-gram analysis and readability scores of annual reports. In *Advances in artificial intelligence, Proceedings of Canadian AI'2009*, pages 39–51. Springer, 2009.

[10] D. J. Cahill. *Lifestyle Market Segmentation*. Hathworth Press, New York, 2006.

[11] Pew Research Center. Twitter news consumers: Young, mobile and educated. http://www.journalism.org/, 2013. Accessed 4 August 2015.

[12] P. Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection*. Springer, Canberra, 2012.

[13] CIHR, NSERC, SSHRC (Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, Social Sciences and Humanities Research Council of Canada. 2010 tri-council policy statement: Ethical conduct for research involving humans. *Tri-Council Policy Statement*, 2010.

[14] Raviv Cohen and Derek Ruths. Classifying political orientation on Twitter: It's not easy! In *ICWSM*, 2013.

[15] R. S. Collica. Customer segmentation and clustering: Using SAS enterprise miner, 2011.

[16] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of Twitter users. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 192–199. IEEE, 2011.

[17] C. Conrad, N. Ali, Q. Gao, and V. Keselj. ELM: An extended logic matching method on record linkage analysis of disparate databases for profiling data mining, 2015. (pending publication).

[18] B. Cooil, L. Aksoy, and T. L. Keiningham. Approaches to customer segmentation. *Journal of Relationship Marketing*, 6(3):9–39, 2007.

[19] V. S. Mookerjee D. Dey and D. Liu. Efficient techniques for online record linkage. *IEEE Transactions on Knowledge and Data Engineering*, 3(23):373–387, 2011.

[20] A. Agrawal D. Palsetia, M. Patwary and A. Choudhary. Excavating social circles via user interests. *Social Network Analysis and Mining*, 22(3):257–282, 2013.

[21] M. Duggan, N. B. Ellison, A. Lenhart, and M. Madden. Social media update 2014. http://www.pewinternet.org/2015/01/09/social-media-update-2014/, 2015.

[22] B. G. Englis and M. R. Soloman. To be and not to be: Lifestyle imagery, reference groups and the clustering of america. *Journal of Advertising*, 24(1):13–28, 1995.

[23] Esri. Esri tapestry segmentation: Methodology. http://downloads.esri.com/esri_content_doc/dbl/us/J9941Tapestry/Segmentation/Methodology.pdf, August 2014. Accessed 18 March 2015.

[24] Esri. Esri demographics. http://doc.arcgis.com/en/esri-demographics/data/data.htm, 2015. Accessed 12 March 2015.

[25] Esri. Esri tapestry segmentation. http://www.esri.com/landing-pages/tapestry, 2015. Accessed 18 March 2015.

[26] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[27] Python Software Foundation. nameparser 0.3.6: A simple python module for parsing human names into their individual components. https://pypi.python.org/pypi/nameparser, 2015.

[28] S. Yang G. Fennell, G. M. Allenby and Y. Edwards. The effectiveness of demographic and psychographic bariables for explaining brand and product category use. *Quantitative Marketing and Economics*, 1:223–244, 2003.

[29] P. Pearce G. Moscardo and A. Morrison. Evaluating different bases for market segmentation. *Journal of Travel & Tourism Marketing*, 10(1):29–49, 2001.

[30] S. Gupta and P. Chintagunta. On using demographic variables to determine segment membership in logit mixture models. *Journal of Marketing Research*, 31(1)(1):128–136, 1994.

[31] K. H. Hahn and E. J. Lee. Effect of psychological closeness on consumer attitutes toward fashion blogs: the moderating effect of fashion leadership and interpersonal lov. *Journal of Global Fashion Marketing*, 5(2):103–121, 2014.

[32] Aaron Hill and Joshua Roesslein. Tweepy: An easy-to-use python library for accessing the twitter api. http://www.tweepy.org, 2015.

[33] Jaakko Hintikka. *Knowledge and belief: an introduction to the logic of the two notions*, volume 181. Cornell University Press Ithaca, 1962.

[34] M. Hofmann and R. K. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. CRC Press, London, 2014.

[35] Chris DL Hunt. Privacy in the common law: A critical appraisal of the Ontario court of appeal's decision in Jones v. Tsige. *Queen's LJ*, 37:665, 2011.

[36] Strategic Business Insights. Vals brochure. http://www.strategicbusinessinsights.com/vals/free/2012-05-VALSbrochure.pdf, 2012. Accessed 18 March 2015.

[37] Strategic Business Insights. The US VALS^{TM} survey. http://www.strategic businessinsights.com/, 2015. Accessed 12 March 2015.

[38] G. Carpenter J. Stockard and L. R. Kahle. Continuity and change in values in midlife: testing the age stability hypothesis. *Experimental Aging Research*, 40:224–244, 2014.

[39] H. Sun J. Wu and Y. Tan. Social media research: A review. *Journal of Systems Science and Systems Engineering*, 22(3):257–282, 2013.

[40] L. R. Kahle. Social values in the eighties: A special issue. *Psychology & Marketing*, 2:231–237, 1985.

[41] L. R. Kahle. Alternative measurement approaches to consumer values: The list of values (LOV) and values and life style (VALS). *Journal of Consumer Research*, 13:405–409, 1986.

[42] Frank R. Kardes, Maria L. Cronley, and Thomas W. Cline. *Consumer Behavior 2e*. CT: Cengage Learning, Sanford, 2015.

[43] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING*, volume 3, pages 255–264, 2003.

[44] A. Koponen. Personality characteristics of purchasers. *Journal of Advertising Research*, 1(1):6–12, 1960.

[45] John Locke. *Two treatises of government*. Cambridge University Press, 1965.

[46] S. Basu M. Bilenko and M. Sahami. Adaptive product normalization: Using online learning for record linkage in comparison shopping. In *Proceedings of the 5th International Conference on Data Mining*, 2005.

[47] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[48] A. Mitchell. *The Nine American Lifestyles: Who We Are and Where We're Going*. Macmillan Pub Co, London, 1983.

[49] V. W. Mitchell. How to identify psychographic segments: Part 1. *Marketing Intelligence & Planning*, 12(7):4–10, 1994.

[50] V. W. Mitchell. How to identify psychographic segments: Part 2. *Marketing Intelligence & Planning*, 12(7):11–17, 1994.

[51] M. Bishop N. Barber, P. J. Kuo and R. G. Jr. Meauring psychographics to assess purchase intention as willingness to pay. *Journal of Consumer Marketing*, 29(4):280–292, 2012.

[52] Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

[53] C. Nie and L. Zepeda. Lifestyle segmentation of us food shoppers to examine organic and local food consumption. *Appetite*, 57:28–37, 2011.

[54] A. Osterwalder and Y. Pigneur. *Business Model Generation*. Self Published, 2010.

[55] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326, 2010.

[56] S. Penaloza. Researching ethnicity and consuption. *Handbook of Qualitative Research Methods in Marketing*, pages 547–559, 2006.

[57] A. Felfernig R. Burke and M. H. Goker. Recommender systems: An overview. *AI Magazine*, 32:13–18, 2011.

[58] D. Delen R. Sharda and E. Turban. *Business Intelligence: A Managerial Perspective on Analytics*. Pearson, New Jersey, 2014.

[59] John Rawls. *A theory of justice*. Harvard university press, 2009.

[60] Equality Rights. Canadian charter of rights and freedoms. *Toronto: Carswell*, 1985.

[61] E. Levy S. Brobst and C. Muzilla. Bi experts' perspective: Enterprise application integration and enterprise information integration. *Business Intelligence Journal*, 10(2):27–32, 2005.

[62] Amartya Sen. *The idea of justice*. Harvard University Press, 2011.

[63] R. Sherman. Back to the basics of data warehousing. *DM Review*, 18(10):36, 2008.

[64] Calvin Thomas, Vlado Kešelj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In *Mechatronics and Automation, 2005 IEEE International Conference*, volume 3, pages 1569–1574. IEEE, 2005.

[65] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[66] A. Camanho V. Migueis and J. F. e. Cunha. Customer data mining for lifestyle segmentation. *Expert Systems with Applications*, 39(10):9359–9366, 2012.

[67] D. B. Valentine and T. L. Powers. Generation y values and lifestyle segments. *Journal of Consumer Marketing*, 30(7):597–606, 2013.

[68] X. Wang and J. Ling. Multiple valued logic approach for matching patient records in multiple databases. *Journal of Biomedical Informatics*, 45:224–230, 2012.

[69] Xiaoyi Wang. *Matching records in Multiple Databases Using a Hybridization of Several Technologies*. PhD thesis, Department of Industrial Engineering, University of Louisville, 2008.

[70] Samuel D Warren and Louis D Brandeis. The right to privacy. *Harvard law review*, pages 193–220, 1890.

[71] A Weinstein. *Market Segmentation: Using Niche Markets to Exploit New Markets*. Probus Publishing, Chicago, 1987.

[72] A. Weinstein. *Handbook of Market Segmentation: Strategic Targeting for Business and Technology Firms*. The Hathworth Press, New York, 2004.

[73] W. D. Wells. *Life Style and Psychographics*. American Marketing Association, USA, 1974.

[74] S. S. Weng and H. L. Chang. Using ontology network analysis for research document recommendation. *Expert Systems with Applications*, 34:1857–1869, 2008.

[75] S. S. Weng and M. J. Liu. Feature-based recommendations for one-to-one marketing. *Expert Systems with Applications*, 26:493–508, 2004.

[76] S. S. Weng and C. Wen-Tien. Using contextual information and multidimensional approach for recommendation. *Expert Systems with Applications*, 36:1268–1279, 2009.

[77] D. R. Wilson. Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage. In *Proceedings of International Joint Conference on Neural Networks*, 2011.

[78] Reid Wilson. The problem with modern polling, in one chart. http://www.washingtonpost.com/blogs/govbeat/wp/2014/03/12/the-problem-with-modern-polling-in-one-chart/, 2014.

[79] W. E. Winkler. Machine learning, information retrieval, and record linkage. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2000.

[80] W. E. Winkler. *Record Linkage Software and Methods for Merging*. U.S. Bureau of the Census, Washington, 2001.

[81] D. Yankelovich and D. Meer. Rediscovering market segmentation. *Harvard Business Review*, 84(2):122–131, 2006.

[82] M. Zanker and M. Jessenitschnig. Case-studies on exploting explicit customer requirements in recommender systems. *User Modeling and User-Adapted Interaction*, 19(1):133–166, 2009.