Predicting Political Donations Using Twitter Hashtags and Character N-Grams*

Unnamed Authors[1]

***Abstract*— We describe a novel approach for predicting political donations and performing psychographic segmentation based on social data linked to election donation records. The role of microblogs in enterprise informatics, specifically in relation to customer relationship systems is highlighted. Algorithms trained on social data can be used to interpret and detect prospects' psychographic information. Contrasted with past approaches which focused exclusively on a single source of social data, the method being presented allows us to use an objective gold standard by linking Twitter and election records. Two experiments were conducted using data collected from 438 Twitter users, half of which are linked with donation event records collected from the United States Federal Election Commission. Probabilistic, entropy and kernel approaches were tested for predictive accuracy, while the CNG technique is explored as an alternative. The CNG algorithm was found to predict political affiliation 17 percentage points above the majority classifier, exceeding benchmarks suggested by the literature. A NaveBayes word n-gram approach was found to outperform CNG at predicting donations by predicting political donations. Insufficient performance and poor reliability of standard word n-gram techniques in opinion detection reveal skepticism about past work on political affiliation analysis from social data alone. This suggests that prospecting systems may benefit from constructing algorithms using data linked to external sources.**

I. INTRODUCTION

The commercial promise of social media is its ability to offer new insight into the activities, interests and opinions (AIO) of consumers. Since its inception as a public "microblog" platform, Twitter has raised considerable interest. Before social media, it was relatively expensive to acquire insight into market segments, as research into AIO required intensive focus group and survey research [1], [2]. Today we might acquire insight into the AIO of prospective customers from automated systems that analyze publicly available data on social platforms, as is characteristic of commercial social insight cloud services. The use of web data in AIO research further benefits from the linkages such systems might perform on other web data.

Motivated by the need for better public understanding of how social marketing platforms might work, this paper describes a process for generating insight into the political opinions and behaviours using publicly available web data. Using records from the United States Federal Election Commission (FEC), labels are generated that represent an individual's recent political giving behaviour and political party affiliation. The labels are then linked to Tweet summaries, which can be used as a "gold standard" for supervised learning. Based on objective data from the FEC, this gold standard represents a novel technique for performing social mining using data from disparate datasets. Four machine learning techniques are tried and assessed on benchmarks from prior political social mining literature. Though more research can be done to refine the approach and test it on a larger scale, we find reasons for raising questions about the validity of mining performed on social data alone.

*This work was not supported by any organization

The paper is organized as follows. Related works are explored in section 2, while Section 3 describes the theoretical framework and experiment design. Section 4 describes the experiments while Section 5 discusses the results and implications of the work. Section 6 concludes the paper.

II. RELATED WORK

Though there is little work on political donor prospecting problems, there is a vast literature on automated customer classification. We describe much of the closely related work below.

A. Web Prospecting and Recommendation Systems

Industrial solutions to segmentation and prospect classification problems are often data-driven and particular to a set of domain-specific problems. Recommendation 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 [3]. Some recommendation systems use supervised learning where a domain expert evaluates the profiles manually, creating labels according to the desired class [4]. In the specific case of social media, supervised learning might be performed on the basis of psychographic profiles related to Activities, Interests or Opinions [1], [2]. This method of using a domain expert to generate classes or ontologies can be applied in a case such as political classification in American politics, where prospects may identify with political parties or exhibit lifestyle characteristics such as willingness to contribute to political giving.

Alternatives to expert-generated classification might include semi-supervised and unsupervised approaches. In semi-supervised approaches, products are recommended to users based on demographic generalizations generated through unsupervised machine learning. The recommendation system specified by Zanker and Jessenitschnig used customer ratings and clickstreams to generate labels for products [5]. Using this method, a system for rating prospects could be generated without the direct engagement of the end users. Weng and

Liu (2003) describe a system for lead generation by clustering customers by interest, rather than the prospects themselves [6]. 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 [7].

B. Record Matching

One of the primary challenges of performing prospect generation from unstructured 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. One approach is to perform record linkage on identifiers such as names, addresses, or date of birth. Rule-based techniques might use phonetic algorithms such as Soundex, dictionaries, or the Levenshtein distance [8] to account for variance 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 [9].

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 or clustering. Probabilities are then compared and are accordingly assigned a confidence rating based on the criteria [10]. Though advanced unsupervised techniques might intuitively seem to perform better, they do not fare much better than rule approaches, and come with drawbacks [11], [14]. Supervised learning might be an ideal alternative, offering greater accuracy with matching trees or probabilistic techniques (eg. Naïve Bayes) [12], [13]. Though these techniques are promising, it exposes an operational limitation on experiments with web data: highly successful matching techniques have a "gold standard" with which to perform supervised learning. 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. Manual matching would be redundant, as humans would simply follow a series of rules or matching criteria. We can conclude that rule-based record linkage might be sufficient for our task, provided that linkages are made on records with both name and location identifiers.

C. Political Opinion Mining on Twitter

Increasingly, web marketers are leveraging social media data from Facebook or Twitter to identify customer trends and characteristics [15]. Using techniques such as sentiment analysis [16], brand mentions and network analysis [17], researchers are able to identify user interests and produce customer insight. Though there have not been any academic attempts to generate data-driven psychographic profiles (eg. "Wealthy Republican-Leaning") 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 [18] 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 [19] 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 [20], 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. [21] 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 with classifying Twitter profiles according to political giving this way [22]. 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. They conclude that many of the challenges are rooted in the incorrect classification of users.

D. 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 stock prediction [24], Alzheimer's detection [25] and other tasks that detect subtle text differences. Originally proposed by Keselj et al. for the author attribution problems [23], CNG might be well suited to the detection of subtle psychological differences in writing styles between individuals exhibiting different psychographic qualities. Where other models using words make predictions based on the semantic 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 fo-

cuses 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.

III. THEORY AND APPROACH

Our task is to test an approach to activities, interests and opinions (AIO) profiling that integrates social media data and behavioural records. The data are matched and integrated using a rule-based matching method, and are then labelled according to features reflective of political affiliation and giving patterns. By integrating the Twitter with the FEC, we can see whether AIO has an impact on actual giving behaviours or political affiliation, rather than relying sentiment or similar analysis from tweets alone. This section describes the theoretical implications of this, as well as the nuances of our experiment design.

A. 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. The data must be summarized, integrated, labeled and stored in a way that achieves our goal.

Industry partners helped us attain data dumps of FEC transactions and Twitter users based on the U.S. State of Missouri. For the purpose of labeling, we took an individual's most recent FEC record since 2008 and discarded the other records. We matched Twitter users with personal identifying information, such as names, addresses and occupations, which were shared among both sets. This of course yields a very small subset of the total population between the two sources. In order to manage the problem of duplicate matches and duplicate Twitter accounts, all Twitter records with duplicate names and addresses are removed. Collecting bulk data in this way allowed us to 219 quality matches.

B. Techniques and Hypotheses

Rather than using an expert approach, we leverage the value of our factual labeled data to test a word unigram technique that does not require manual labeling. Theoretically, this gets us closer to the "ground truth" of peoples' opinions. In our experiments, we test four predictive models. The first three are standard machine learning tools: C4.5, SVM and NaiveBayes, which use the hashtag n-grams. By creating a dataset which summarizes the hashtag word unigram mentions, we might assess the relationship between mentions and political giving, similar to Conover et al. [21]. Given that this has been tried, we can evaluate our success against Cohen et al. [22] who find that it is challenging to make predictions about political affiliation with accuracy over 65%. In addition, Previous explorations with character n-grams have been successfully utilized in authorship attribution [23] and financial forecasting [24] and could theoretically be applied to detect unforseen relationships between authors who have political affiliations or are likely to make political donations. In addition to the aforementioned solutions, we explore the application of the CNG technique, which compares the results from multiple character n-grams, up to 10.

For each experiment, we test the four techniques at a specific task. The first experiment involves predicting affiliation using hashtag word unigrams, along the lines of Conover et al. The second experiment involves predicting donation propensity. We can form a series of hypotheses about the relationship between the classifiers and the techniques:

- H1 Probabilistic models can predict political affiliation with 65% accuracy, along the findings of Cohen et al.
- H2 CNG will outperform the best probabilistic models in both affiliation and donation tasks.
- H3 Machine learning will be less apt to build models on Affiliation than Donations.

IV. EXPERIMENTS

In this section, we describe the data, preparation, procedure and experiments undertaken to identify political affiliation and political donation propensity.

A. 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. In addition to the FEC filings, an industry partner provided 119 071 Twitter records containing user names and profile summaries from the US state of Missouri. Using the name features of the FEC and Twitter Profiles, we were able to perform linkages on first name, last name and city fields. Each profile was then manually matched based on occupation criteria provided in both the Donor and Twitter records.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.

The next phase involves extracting and processing user tweets for analysis. The Twitter API is used to extract Tweets. Using the Tweepy library [26], we are able to construct a Python program that reads a list of Twitter users, calls the Twitter API to collect Tweets for that user, and records the tweets in CSV format. In addition, to the 219 matched records, 219 other Missouri profiles were randomly selected to serve as a control for political donation detection. Each of these profiles were manually investigated to ensure that they were not spam accounts. 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.

B. Experiment 1: Predicting Political Affiliations

The Affiliations data consists of the tweet records from 219 individuals for whom we have FEC matches. All predictive techniques utilized in this experiment were classifiers, 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 87 "Unknown" profiles, extracted from the party of the most recent donation.

In order to determine the optimal number of features, for the techniques that utilized word n-grams, the 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.

For the CNG for the task we utilized the hashtag collection'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. 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 training folder, while 73 files were included for testing. 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.

C. Experiment 2: Predicting Donations

Much like the experiments with Affiliations, Donation experiments are conducted by applying the techniques to collections of hashtag word 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 made a political donation since 2008. 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.

As with the previous experiment, 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 yielding 438 total instances. 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. As 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. CNG was also applied as described above, with 292 profiles being used for training and 146 for testing.

V. RESULTS AND DISCUSSION

A. Results: Experiment 1

Hashtag word unigrams, which performed well with Conover et al., did not perform well on the affiliation dataset. The best classifier using aggregated hashtag word unigrams achieved 42% accuracy, just three basis points above the majority classifier. These results were initially disappointing. We were largely unable to repeat the findings of Conover or Cohen, which indicated that we might be able to attain accuracy at least 15% above the majority classifier. We thought that the error might be attributed to the noise generated by the UNK class, which was absent from the Conover set. However, results from CNG were surprising. CNG was able to build a model with 55% accuracy using character the 500 most common bigrams, and another with 55% accuracy using the 1000 most common character trigrams. The results of CNG are described in Figure 1, where N represents the number of characters in the n-gram (eg. trigrams, 10grams) and L represents the number of n-grams used in the classification algorithm.

These results are 16 percentage points above the majority classifier, and are reflective of a degree of predictive accuracy, albeit not a high degree of accuracy, and are roughly in line with Cohen's criticism. CNG was thus able to exceed C4.5 by 13 percentage points, and we were largely unable to replicate the findings of Conover et al, which asserted the predictive power of the Naivebayes algorithm at political mining tasks.

B. Results: Experiment 2

Techniques that utilized Word unigrams performed comparatively better at predicting donations, even exceeding the performance of CNG. The best classification technique

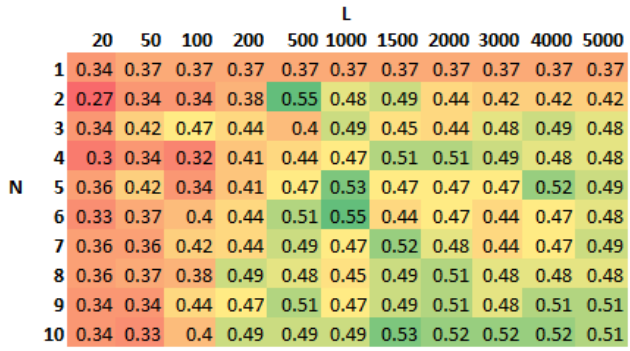

Fig. 1. Results for CNG Political Affiliation on Words

was NaiveBayes, with a predictive accuracy of 66.2 % and 3368 attributes. This is also 16 percentage points above the majority classifier. The results are summarized in Figure 3.

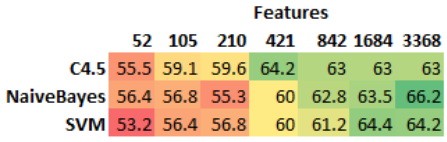

Fig. 2. Results of Algorithms at Donation Propensity Tasks

Though CNG did produce predictive results, it did not exceed the performance of NaiveBayes at this task. CNG managed to produce model with 61% predictive accuracy using the 1000 most common 9-grams. The fact that our data yielded weak results on donation predictions further cemented confidence that our data, though noisy, held some predictive potential.

C. Discussion

If we revisit the hypotheses outlined at the outset of the experiment, we realize that the evidence interesting, though by no means conclusive results. Our experiments raised some skepticism about H1, the hypothesis that we could replicate the results described in Conover et al. [21] and Cohen et al [22]. Given the failure of hashtag word unigrams to have any significant impact on political donation affiliation, and yet to derive some predictive power from the data, we must question the effectiveness of political mining on Twitter alone. If political opinion mining reflected actual political affiliation of most Twitter users (as opposed to merely the vocal minority), replicating Conover's techniques should have yeilded results at least in line with CNG. The problem was likely not with our methodology, as CNG was able to generate predictions with a degree of accuracy. In line with the criticism of Conover in Cohen et al., we must further investigate the methodology behind political mining on Twitter.

Though CNG exceeded the performance of other algorithms at political affiliation tasks, probabilistic and SVM models performed better at donation tasks. We speculate that this may be because the features that impact donations have less to do with subtle personality traits, and more to do with overt traits such as wealth or use of political activist hashtags. There is evidence to believe that CNG does not perform better than other methods at all tasks, and that its application is in detection of subtle personality traits. We must raise H2 into question as it likely overstates CNG's claim.

Concerning H3, it would seem that our method of labeling Tweets using FEC data facilitates the construction of predictive models that might otherwise be inaccurate. If we dig deeper, the models that we applied to this dataset may potentially offer insight that would not otherwise be available to us. The most successful n-gram model at affiliation prediction 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 1 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 I
TEN POPULAR BI-GRAMS FROM EACH CLASS

Though the DEM and REP n-grams share most of the common bi-grams, there are some differences in the distance between the frequency of these bi-grams popular n-grams. DEM, for instance, seem more likely to use the "_ @" n-gram, which is indicative of a mention of another Twitter user. The further we go with the n-grams, there is much more variance in the frequency between the grams, especially around the 500 mark. With a larger dataset, we might do further investigation on the social differences between DEM and REP using the differences in frequency of the grams. 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 potentially extract marketing insight about the users.

Perhaps the strongest feature of our 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.

D. Limitations

The method explained in this paper was limited in scope, specifically to political giving, specifically to a very narrow set of 219–438 users. The data sample size is too small to draw strong conclusions about the hypotheses. However, future work with a larger dataset using this technique might yield very interesting results. Importantly, however these trends and findings in political giving might not translate into the wider world of psychographic profiling. Additionally, 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. 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 with certainty whether the affiliation classifier performs better than those described in the literature. However, 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.

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. Future research might involve self-identified Twitter users on a larger scale, which might solve the problem of record linkage.

VI. Conclusion and Future Work

The use of publicly available 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, lending support to Cohen's criticisms and the skepticism about hashtag word unigrams to produce comparable results to CNG suggests that there is considerable potential for the application of this gold standard and related data mining techniques to affiliation problems. With cleaner data, other advanced methods, such as neural networks, might also yield robust results when applied to psychographic research using this type of gold standard.

The profiling system is also highly scalable. The system used in these examples used two labels to create profiles: political affinity and political donation records. Additional features might be investigated in the same vein. Examples might include using other disparate data to create labels about hobbies or interests. The classification algorithms built from these techniques can then be applied to create more comprehensive social media informatics and customer prospecting systems. 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. The promising results of the novel method for dataset labeling and the application of CNG to political affiliation could offer a new area of inquiry.

References

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

[2] 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)

[3] R. Burke, A. Felfernig and M. H. Goker: Recommender Systems: An Overview. AI Magazine 32, 13–18 (2011)

[4] G. Adomavicius and A. Tuzhilin: Using Data Mining Methods to Build Customer Profiles IEEE Computer, February, (2001).

[5] 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)

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

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

[8] Navarro, Gonzalo: A guided tour to approximate string matching. ACM computing surveys (CSUR),33: 1 (2001)

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

[10] Xiaoyi Wang: Matching records in Multiple Databases Using a Hybridization of Several Technologies. Department of Industrial Engineering, University of Louisville (2008)

[11] W. E. Winkler: Machine Learning, Information Retrieval, and Record Linkage. Proceedings of the Survey Research Methods Section, American Statistical Association (2000)

[12] D. Dey, V. S. Mookerjee and D. Liu: Efficient Techniques for Online Record Linkage. IEEE Transactions on Knowledge and Data Engineering, 3:23, 373–387 (2011)

[13] D. R. Wilson: Beyond Probabilistic Record Linkage: Using Neural Networks and Complex Features to Improve Genealogical Record Linkage. Proceedings of International Joint Conference on Neural Networks (2011)

[14] M. Bilenko, S. Basu and M. Sahami: Adaptive Product Normalization: Using Online Learning for Record Linkage in Comparison Shopping Proceedings of the 5th International Conference on Data Mining (2005)

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

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

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

[18] Alexander Pak and Patrick Paroubek: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC, 10, 1320–1326 (2010)

[19] Esuli, Andrea and Sebastiani, Fabrizio: Sentiwordnet: A publicly available lexical resource for opinion mining. Proceedings of LREC, 6, 417–422 (2006)

[20] Tumasjan, Andranik and Sprenger, Timm Oliver and Sandner, Philipp G and Welpe, Isabell M: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178–185 (2010)

[21] Conover, Michael D and Gonçalves, Bruno and Ratkiewicz, Jacob and Flammini, Alessandro and Menczer, Filippo: Predicting the political alignment of Twitter users. Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing

(SocialCom), 2011 IEEE Third International Conference on, 192–199, (2011)

[22] Cohen, Raviv and Ruths, Derek: Classifying Political Orientation on Twitter: It's Not Easy! ICWSM (2013)

[23] Kešelj, Vlado and Peng, Fuchun and Cercone, Nick and Thomas, Calvin: N-gram-based author profiles for authorship attribution Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING, 3, 255–264 (2003)

[24] Butler, Matthew and Kešelj, Vlado: Financial forecasting using character n-gram analysis and readability scores of annual reports. Advances in artificial intelligence, Proceedings of Canadian AI'2009, 39–51, (2009)

[25] Thomas, Calvin and Kešelj, Vlado and Cercone, Nick and Rockwood, Kenneth and Asp, Elissa: Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. Mechatronics and Automation, 2005 IEEE International Conference, 3, 1569–1574 (2005)

[26] Hill, Aaron and Roesslein, Joshua: Tweepy: An easy-to-use Python library for accessing the Twitter API (2015)