Mining Social Media Data from Sparse Text: An Application to Diplomacy

Cecil Chua[1], Xiaolin (David) Li[1], Mala Kaul[2], and Veda C. Storey[3]

[1] Information Systems and Operations Management Department, University of Auckland Business School,
University of Auckland, Auckland 1010
aeh.chua@auckland.ac.nz, david.xiaolin.li@gmail.com
[2] Dept of Information Systems, College of Business, University of Nevada,
mkaul@unr.edu
[3] Department of Computer Information Systems, J. Mack Robinson College of Business
Georgia State University, Box 4015 Atlanta, GA 30302
vstorey@gsu.edu

Abstract. Publicly available data in social media provides a wealth of unstructured data for applications, such as sentiment analysis and location-based services. This research analyzes a specific application of diplomats, who seek to understand the people with whom they must negotiate. Social media data about a negotiating partner can, potentially, be used to build a profile of that partner. However, such data is difficult to mine effectively because it has sparse text with high dimensionality. This research uses a design science approach to develop a method for extracting critical information from sparse text. The method mines sparse text from publically available Facebook data to extract patterns from individual communications. The method is applied to Facebook posts of a political figure to identify meaningful categories of information for insightful inferences. Preliminary evaluation shows support for the method.

Keywords. Design application, diplomat, natural language processing, sentence clustering, sparse text mining, semantic chain

1. Introduction

Intelligence, the collection of information of value, is important to organizations and nations across the world. Diplomats, for example, often desire a deep understanding of the people with whom they must negotiate. This includes knowing who else the other party has talked to, when, where, and on what topic. Much intelligence is gathered from public or near-public sources. Social media has become a ubiquitous source of publically available information, which can be found in one place and applicable in many domains, including politics and foreign affairs, enabling one to derive "actionable information" [13]. Such information is valuable prior to engaging in discussions or negotiations. Manual review and analysis of such data, however, is time consuming, so a (semi-) automated approach to analyzing the data could significantly reduce the time spent on dealing with the large volume of data and might yield patterns, hints, or indicators of useful information. For example, the following post from a Prime Minister deals with crime and, one could infer, the results of an election.

"People want to know they are safe on their streets and in their homes - and our plan to make sure they are is working. We've cut red tape and given police one simple target: cut crime. And, thanks to the efforts of hardworking police officers, crime is down by over 10% since the election."

Existing text-mining algorithms require a large text corpus[1] for extracting useful information. Data on social media such as Facebook, however, tends to be sparse text, where each data point can comprise 10 words or less, whereas most text-mining algorithms require each data point to have at least 70 words. The research problem then becomes how to derive meaning from inferred semantic relationships between words, focusing on fragmented text in social media. This is an important design problem, common in many domains. Many businesses use social media platforms to receive customer complaints and/or feedback. Businesses would like to aggregate these complaints to identify systematic issues to correct. For example, KLM, the flag carrier airline of the Netherlands, is reported to answer 92% of complaints on their Facebook page with an average first response time in less than 30 minutes [9]. They currently do so manually. (Semi-) automated processing could potentially reduce the response time and labor cost. Similarly, businesses can mine the social media pages of their competitors to infer competitor strategy.

The objective of this research is to develop a method to mine (cluster) large amounts of sparse text found in social media data. The problem arises from: (1) the number of documents in the corpus being very large; and (2) the length of each document being very short. To evaluate the method, we apply it to Facebook data of politicians and/or public figures to derive actionable intelligence. The contribution is to provide a method for clustering vast amounts of social media data, with a sparse sentence structure with no explicit embedded metadata (e.g., hashtags), to develop a partial profile of an individual as might be used for negotiation purposes. A design science research approach is applied to develop the method artifact.

2. Related Research

Existing benchmarks for clustering "sparse text" are considerably different from those pertaining to social media data. Benchmarks of sparse corpii designate content of 6-7 paragraphs and approximately 1000 words.[2] Even research on "short sentence clustering" within narrow domains (e.g., medical corpus), includes approximately 70 words [2]. Research on information retrieval considers documents with less than 60 words as "short" [7]. Some research has been applied to Twitter [14], but such research relies on structured elements of Twitter communication such as #hashtags. The Facebook corpus has on average, 11 distinct words, after excluding "stop words," etc. An associated problem with sparse text is high dimensionality; that is, there are many possible ways in which a naïve algorithm can group the few words across sentences. Thus, to solve the sparse text problem, it is important to address the challenges of short sentences and high dimensionality.

[1] The set of sentences to be processed is referred to as a corpus.

[2] http://about.reuters.com/researchandstandards/corpus/statistics/index.asp

Prior approaches have incorporated information from external sources to enhance naïve algorithms. The most frequently employed algorithms are based on either TF-IDF (i.e., term frequency – inverse document frequency), or bag of word representations [6]. TF-IDF counts the number of times words appear within and across documents. Words that appear frequently in a few documents are granted higher weight than words appearing frequently across many documents or words that appear infrequently. Bag of word representations create a two-dimensional matrix of the words and attempt to calculate similarity scores between them.

In addition to this frequency-based approach, there are knowledge based approaches where information is captured in the form of a lexicon (e.g., WordNet) or an ontology, or even an information source containing links from which an algorithm can infer information, such as a wiki. Supervised learning approaches also exist. None of the approaches, however, are designed to address the sparse text problem, such as that found in Facebook and tend to produce suboptimal results. Nevertheless, these approaches are a promising initial basis for our research.

3. Research Approach

This research follows a design science approach [8] to propose a method for mining sparse text and identifying patterns. The method is instantiated to provide proof-of-concept by applying the data mining method to three months of Facebook data of a politician. In general, the proof-of-concept stage involves instantiating a concept to provide evidence that the system can perform as conceptualized, to demonstrate feasibility [11]. Technical, observational, empirical, and theoretical insights together generate a potentially unique or innovative solution. The nominal process sequence [12] followed in this research is summarized in Table 1.

Table 1: Design Science Approach to Sparse Text

Step	Sparse Text Sentences
Problem identification and motivation	Valuable information can be obtained from text mining of social media data. How can valuable insights be extracted when text is sparse?
Objectives of a solution	Develop a method (artifact) for sparse text mining
Design and development	Method based upon data mining techniques, natural language parsing, and semantic clustering.
Demonstration	Application to Facebook data over 3 months
Evaluation	Application to corpus
Communication	Document analysis of resulting chains

For the initial development, the input to the method is a corpus of Facebook posts, e.g.:

> *"Our long-term economic plan is helping people across the country who want to work hard and get on in life …"*

"The biggest quarterly increase in employment on record. More jobs means more security, peace of mind and opportunity.."
"We need everyone in the country to get behind our long-term economic plan. Our plan builds a stronger, more competitive economy and secures a better future…"

The focus of this paper is extracting information that specifically addresses "what" a politician is talking about. Eventually, the method should be able to extract information that also addresses "who," "what," "when," "where," and "how."

4. Method for Sparse Text Mining

This research is based upon word and phrase parsing as well as clustering algorithms. The main steps are: *preprocessing*, *similarity calculation*, *generation of semantic chains*, and *sentence clustering*. The fundamental idea behind the semantic clustering is that, with the help of WordNet, the similarities among all of the words that appear in a corpus are compared. The similar words form different clusters, based upon the similarity scores calculated, are called "semantic chains." Then, all of the posts are grouped around the "selected" semantic chains of median length.

Preprocessing. The text is preprocessed by: (1) removing stopwords, (2) word lemmatization, and (3) indexing the words. Similar processes have been implemented in systems such as the SMART Information Retrieval System [4] and the Snowball text-mining system [1].

Stopword Removal. As a classic pre-processing strategy [5], stopword removal is used for two reasons. First, stopwords occur frequently so eliminating them greatly speeds up text mining algorithms. Second, these words disrupt many text mining algorithms because they rely on matching common words across pieces of text. There are two kinds of stopwords: (1) common (e.g., "the", "and," "in"); and 2) domain-specific. In the diplomat domain, for example, continent-words such as "Asia" and "Europe" are not helpful when one wants to know which specific countries in Asia and Europe the prime minister of New Zealand**,** John Key**,** visits. This research adopts the Snowball stopword list, comprised of 102 words [1]. The adoption of the Snowball list resulted from an earlier empirical test that compared Snowball to other stopword lists in our data set.

Word Lemmatization. Lemmatization removes inflectional endings. For example, in simple lemmatizers, the words "organize," "organized," and "organizing" are treated as the same word. We employ the Lemmatizer in NLTK [3]. This lemmatizer derives the root of a word, but then identifies all words in WordNet that have that word as a root. This is necessary, because we employ WordNet later to perform similarity calculations.

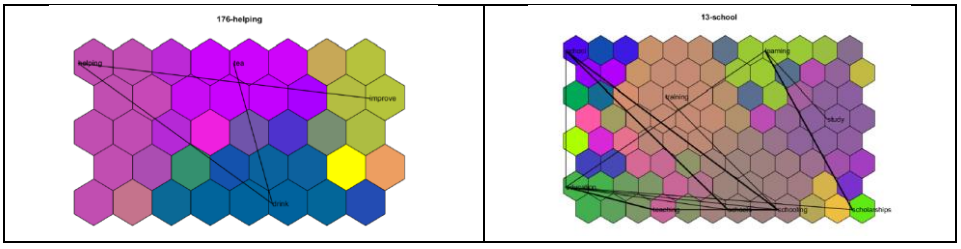
Indexing. Each remaining unique word is given an index number, with their frequencies of appearance throughout the entire corpus recorded. The frequency of each word is an important base statistic used in text mining [6].

Similarity Calculation. The similarity calculation is based on WordNet [10]. In WordNet, a hyponym is a word or phrase that is "similar to" (semantically) another word. A similarity score between every possible pair of words is calculated.

Generation of Semantic Chains. Semantic chains have been used extensively for keywords extraction. A semantic chain is a sequence of words (e.g., education, budget) deemed to be related. We generate semantic chains incrementally. First, we generate all semantic chains comprising two words. We then attempt to add a third word to each of these chains, etc. The size and quantity of chains is determined based upon a similarity threshold. The longer and more chains we have, the more complex the later part of the processing will be. A threshold of 0.4 appears to produce computationally tractable chains. Chains that are "too long" are eliminated, as well as those that are "too short." Preliminary analysis shows that chains of approximately 5-10 words seem to be best, which is what is currently implemented.

Besides selecting median-sized chains, chains are eliminated by calculating a normalized connectivity score of a chain, defined as $\frac{S}{n \times (n-1)}$, where S is the sum of the similarity scores of each pair of words and n is the number of words in the chain. This determines the "average" similarity score across all words in a chain. A more cohesive chain is expected to have a higher normalized connectivity score, as can be observed through visualizing the detailed structure for each semantic chain.

To illustrate, for our sample corpus, chain 176 (helping) in Figure 1a, has a normalized connectivity score of 0.75. The words in the chain do not form a cohesive topic; e.g. the words "tea" and "helping", "drink" and "improve" are not related semantically. On the other hand, in Figure 1b, where the normalized connectivity score is 1.44, the words in this semantic chain 13 (school) are gathered around the topic "education," and are more semantically related. The 150 semantic chains with the highest normalized connectivity score are used for the analysis. Otherwise, the number of semantic chains would be too many to compute effectively.



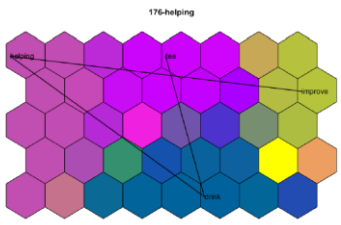


Figure 1a. Connectivity Graph for chain 176 (normalized connectivity score 0.75)

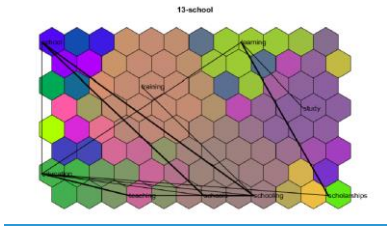


Figure 1b. Connectivity graph for chain 13 (normalized connectivity score 1.44)

Figure 1: Sample Connectivity graphs for 2 chains

Figures 1a and 1b are produced by Self-Organizing Maps (SOM) using the term-document matrix generated from the corpus. In a corpus containing d terms and n documents, the term-document matrix is a d *x* n matrix, in which the (i, j)th entry is the frequency of the ith term in the jth document. For the semantic chain i with words denoted by $\{v_1, \cdots, v_{N_i}\}$, the visualization is produced by taking rows in the term-document matrix with index $v_1, \cdots, v_{N_i}$ to train the SOM. After the training, each grid on the SOM is accompanied with an n-dimensional vector. The colors on each grid are produced by first reducing each n-dimensional vector to 2-dimensional using Curvilinear Component Analysis (CCA) and then converting the 2-dimensional vector into color. One can consider the background color on each grid as the similarity derived from the text contexts of the original corpus. After adding the edges by referring to the similarity produced by WordNet, the additional similarity identified by using WordNet is clearer.

Clustering. Any sentence containing a word in the semantic chain is considered to belong to the same "group."

5. Application of Method

The method is being applied to the Facebook page of John Key (Prime Minister of New Zealand) to build, evaluate, and refine it. The corpus consists of 499 extracted posts over a three month period. The resulting 150 clusters that formed were sorted into three groups: 1) "clusters that look reasonable to a human;" 2) "clusters that do not make sense to a human;" and 3) "clusters that somewhat look reasonable to a human." One of the main issues causing sentences to fall into the "do not make sense" category is when clusters mainly comprise verb or action roots, as opposed to sentences in the "look reasonable" category that do not tend to be verbs. The next iteration will strip all such words before clustering to assess whether doing so improves accuracy. There were also words common across many documents in the clusters that do not make sense, implying the need to incorporate TF-IDF into the method. An example of a resulting chain and its evaluation is given in Figure 2.

Chain 92 has the following words:
enjoy loving employ commit enjoyed enjoying
Chain 92 has the following posts around it:
15: Blenheim really turned it on today. Enjoyed meeting so many people during the Royal visit.
26: Enjoyed helping out at Henderson Primary Breakfast Club this morning. 25,000 children each week in 372 schools are taking part in KickStart. It really shows how communities, businesses, and the government can work together to help children in need.
33: Really enjoyed the NZ Shearing Champs this evening in Te Kuiti. Brilliant shearing skills. Congrats to all the winners tonight.
51: On Saturday I enjoyed mixing with the crowds at the Pasifika Festival.
69: Plenty of selfies at O-week at Vic. Loving their enthusiasm.
97: Am enjoying chatting with you all on Newstalk ZB Wellington this morning.
107: On 11 February, I enjoyed taking part in Chinese New Year celebrations at Parliament.
128: At Red Stag timber in Rotorua with Todd McClay MP – it's a local success story, employing 360 staff.
Comments: sentences centre around the word "enjoy". Score 1.67
Rank: Good.

Figure 2: Chain 92 words and posts

As can be seen, the chain scored reasonably well suggesting evidence of a successful application.

6. Conclusion

The mining of sparse text is a general problem for extracting value from social media data. This research proposed a method for addressing sparse sentence structure (sparse text) problems and applied it to the analysis of diplomatic relationships as found in an online presence. A design science research approach was used to create the method artifact through adoption of prior work on natural language parsing and semantic clustering, extended and refined through an iterative process of testing and use. This method is illustrated through the mining of Facebook data to make inferences intended to lead to a profile of a public figure. To test the feasibility of the research, the method has been applied to 499 posts from the Facebook posts of one diplomat. Initial results suggest that the method appears feasible.

The research extends prior work on sparse text mining by providing a method for mining and clustering sparse text resulting in a visualization of patterns using connectivity graphs and chains. This research contributes to social media intelligence by developing and implementing a method for finding patterns in social media data. This general method could potentially be applied to multiple applications of mining sparse text for the purpose of drawing inferences from patterns of connections between words. For example, it could be used for profile development which has applications

in law enforcement, marketing (customer profiles), or other behavior analysis such as in sentiment analysis or in fraud detection. Future research will involve further development of the method and testing and extending it to inferences dealing specifically with questions of "who, what, where, and when" pertaining to diplomats. It will then be tested on other applications.

References

1. Agichtein, E., Gravano, L. (2000) Snowball: Extracting Relations from Large Plain-Text Collections. Proceedings of the Fifth ACM International Conference on Digital Libraries, pp. 85-94. San Antonio, TX: Association for Computing Machinery, 2000.
2. Avendano, D.E.P. (2007) Analysis of narrow-domain short texts clustering. (Diploma of Advanced Studies thesis), Retrieved from http://users.dsic.upv.es/~prosso/resources/PintoDEA.pdf.
3. Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.
4. Buckley, C. (1985) Implementation of the SMART Information Retrieval System. Technical Report, Cornell University Ithaca, NY, USA.
5. Callan, J.P., Lu, Z., Croft, W.B. (1995) Searching distributed collections with inference networks. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21-28.
6. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. (1990) "Indexing by Latent Semantic Analysis," *Journal of the American society for information science* (41:6), pp. 391-407.
7. Jing L., Ng M. K., Yang X., and Huang Z., A Text Clustering System based on k-means Type Subspace Clustering and Ontology, International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol. 2, No. 4, 2008.
8. March, S. T., and Smith, G. F. 1995. "Design and Natural Science Research on Information Technology," *Decision Support Systems* (15:4), pp. 251-266.
9. McCrea, L., (2012) 16 Brands Leading the Way with Exemplary Social Media Customer Service. Ignite Social Media blog.
10. Miller, G., and Fellbaum, C. 1998. "Wordnet: An Electronic Lexical Database." MIT Press Cambridge.
11. Nunamaker Jr, J.F., and Briggs, R.O. 2011. "Toward a Broader Vision for Information Systems," *ACM Transactions on Management Information Systems (TMIS)* (2:4), p. 20.
12. Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3), pp. 45-77.
13. Zeng, D., Chen, H., Lusch, R., Li, S. H. (2010). Social media analytics and intelligence. *Intelligent Systems*, IEEE, 25(6), 13-16.
14. Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110 (15), 5802-5805.