Discovering Context: Classifying Tweets through a Semantic Transform Based on Wikipedia

Yegin Genc, Yasuaki Sakamoto, and Jeffrey V. Nickerson

Center for Decision Technologies, Stevens Institute of Technology
Castle Point on Hudson, Hoboken, NJ 07030 USA
{ygenc,ysakamot,jnickerson}@stevens.edu

Abstract. By mapping messages into a large context, we can compute the distances between them, and then classify them. We test this conjecture on Twitter messages: Messages are mapped onto their most similar Wikipedia pages, and the distances between pages are used as a proxy for the distances between messages. This technique yields more accurate classification of a set of Twitter messages than alternative techniques using string edit distance and latent semantic analysis.

Keywords: Text classification, Wikipedia, semantics, context, cognition, latent semantic analysis.

1 Introduction

Humans are experts in recognizing new and useful messages while ignoring others. They do this by extracting meaning from messages, categorizing messages with related meaning into the same topics, and noticing information that does not fit any existing categories. Attempts to automate this fundamental ability of cognition using semantic models still leave room for improvement (e.g. [1]).

We study how we can categorize messages streaming through Twitter. These messages, called tweets, come in at a rate of more than 600 a second [2], and are often cryptic. If we can find ways of categorizing messages and recognizing new and useful topics in this noisy environment, we may provide automated tools with pragmatic uses: Twitter functions as a large sensor system, and can increase our awareness of our surroundings (e.g. [2-5]).

In an organizational context, the act of decrypting confusing or novel information is called sense-making, and is accomplished in part by asking other people what they think [6]. In artificial intelligence, this act is related to the understanding of context. For example, tweets might be looked up in Wikipedia, and the closest entry to a tweet found [7]. This technique is a promising instance of a larger idea, in which machine algorithms inform themselves by seeking out contextual information [8-12].

In the present paper, we apply several semantic models to tweet categorization. We introduce a Wikipedia-based classification technique. Extending the insight of Michelson and Macskassy [7], we develop a technique for calculating semantic distances between messages based on the distances between their closest Wikipedia

D.D. Schmorrow and C.M. Fidopiastis (Eds.): FAC 2011, HCII 2011, LNAI 6780, pp. 484–492, 2011.

pages: in effect, we regard Wikipedia as a transform space in which measurements can be made. We next describe this classification technique, and two other techniques we will use for comparison.

2 Classification Techniques

Classifying tweets is not an easy task because statistical methods of text classification have difficulty on short texts [13]. Moreover, if emerging topics of conversation are regarded as signal, the vast majority of tweets would be characterized as noise.

Some past studies of tweet classification have examined the use of specific features, such as emoticons [14] and author profiles [15], in improving the classification performance. Other studies have regarded tweets as a window into customer perception [16]; then, the challenge becomes recognizing sentiment. In contrast to these past work, we are interested in categorizing tweets in order to detect topics, which requires the ability to cluster tweets without a priori knowing which features will be important.

Recent work on extracting topics from short texts relies on knowledge bases to find context that is not in the texts. For example, Stone et al. used Wikipedia as training corpus to improve the ability of statistical methods to discover meanings of short texts [17]. Similarly, Gabrilovich and Markovitch used concepts derived from Wikipedia to identify the semantic relatedness of texts [9]. Wikipedia was also used by Michelson and Macskassy [7]: since we build on their model we will discuss it in the next section, after giving an overview of our own process.

Tweets are classified in three steps. First, between-tweet distances are calculated using one of the techniques described next. We map the tweets onto two-dimensional planes using multidimensional scaling (MDS) of the between-tweet distances. MDS helps us interpret the underlying relationships in the data, by allowing us to visually examine the clustering of tweets, similarity between clusters, and the size and internal consistency of the clusters. Then we use discriminant function analysis to measure how well each technique can predict the category memberships of the tweets [18]. Discriminant function analysis predicts a categorical dependent variable, in this case the tweet category, by one or more independent variables, in this case the two-dimensional MDS solution (i.e., the x and y coordinates).

There are many ways of building a between-tweet distance matrix. We picked String Edit Distance [19] and Latent Semantic Analysis [20] to compare with the Wikipedia-based algorithm presented next.

2.1 A Semantic Transform Using Wikipedia

Michelson and Macskassy have developed a model that discovers topics of interests of Twitter users based on their tweets: capitalized non-stop words in tweets are linked to Wikipedia pages; then, topics are derived from socially tagged categories listed in the linked Wikipedia pages [7]. Specifically, topics are discovered by traversing the tree structure of the taxonomy of Wikipedia. We apply a similar technique to a different end. We are interested not in determining the topic of a particular user's set of posts, but instead, understanding topic emergence across many users (e. g. [3, 21]).

In order to do so, we will create a distance matrix between tweets. As shown in Fig. 1, there are two stages: (1) we map tweets to Wikipedia pages, and then (2) compute the distance between the Wikipedia pages as a measure of semantic distance between the tweets. More formally, we regard Wikipedia as a transform space, in which we measure the between-tweet distances:

$$d(message_1, message_2) \propto d(T(message_1), T(message_2)),$$

where d is a measure of semantic distance, and T is a transformation function mapping a message to a page in Wikipedia. The transformation is worth performing for two reasons. First, humans categorize Wikipedia pages based on their meanings, and thus the Wikipedia networks likely reflect semantic networks in human brains. This can be helpful because we are dealing with tweets, short texts humans write online. Second, Wikipedia pages are mapped to categories that are named by the crowd. These categories can serve as topics, eliminating the problem of inferring the meaning of latent topics in Latent Semantic Analysis and other statistical methods.

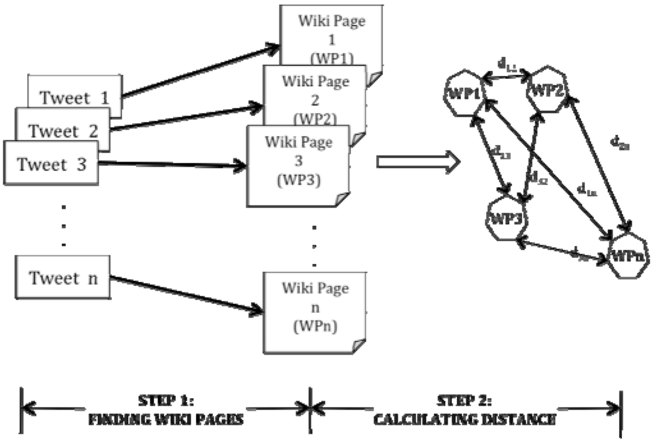

Fig. 1. The two steps involved in calculating distances between tweets using Wikipedia. We use the distance between the two associated Wikipedia pages as an indicator of the distance between the two tweets.

Finding Associated Wikipedia Pages. To associate a tweet to a Wikipedia page, we first identify a set of words for this tweet. The word set includes all the words in the tweet after eliminating certain words in the English stop-words list provided in the LSA package for R [23]. For each word, we check to see if there is a direct page dedicated to the word, and if there is a disambiguation page. The disambiguation page provides a precise mapping to the right page, leading to more accurate distance measures. Then a list of candidate pages for the tweet is found by aggregating each page associated with each word of the word set. We compute a score for each candidate page by counting the number of occurrences of the words in the word set. The page with the highest score is selected as the associated Wikipedia page for the tweet. This process is visualized in Fig. 2.

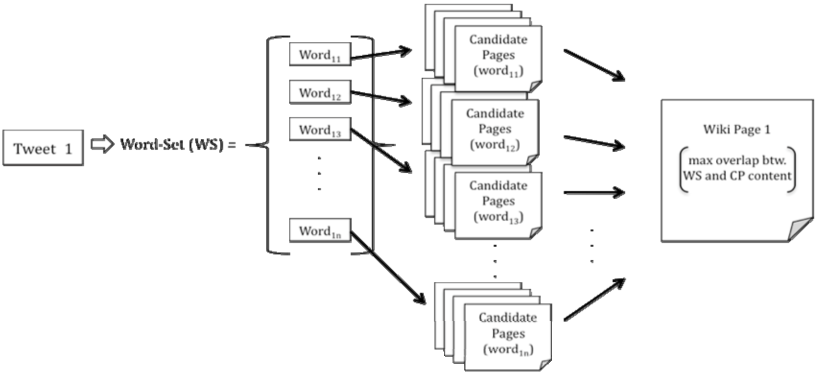
Tweet 1
Word-Set (WS) =
Word$_{11}$
Word$_{12}$
Word$_{13}$
Word$_{1n}$
Candidate Pages (word$_{11}$)
Candidate Pages (word$_{12}$)
Candidate Pages (word$_{13}$)
Candidate Pages (word$_{1n}$)
Wiki Page 1
max overlap btw. WS and CP content

Fig. 2. Finding a Wikipedia page associated with a tweet

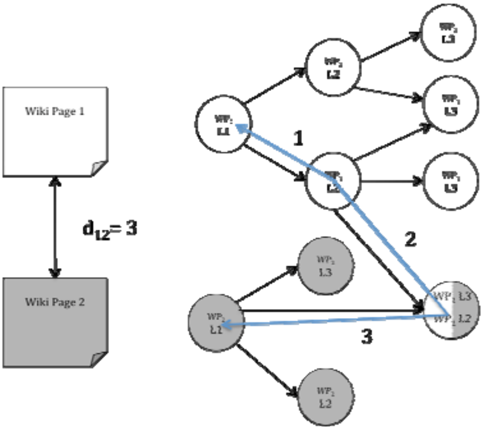

Fig. 3. Calculating the distance between two Wikipedia pages

Calculating Distances. The distance of two Wikipedia pages is calculated based on the link between the categories associates with these two pages. Categories of the Wikipedia pages are linked to one another in a graph structure. A category can be linked to multiple parent categories. We capture the network structure of categories for each Wikipedia page for five levels. We compute the semantic distance between the two Wikipedia pages by finding the length of the shortest path from a category of one page to a category of the other page. Fig. 3 shows an example.

2.2 String Edit Distance

The String Edit Distance (SED) method may work given that tweets are short: related tweets might contain the same set of keywords. In the SED method, the distance between two tweets is found by calculating the number of edits it takes to transfer one tweet to another, also called Levenshtein distance. As an example, the Levenshtein distance between “kitten” and “sitting” is 3 since (1) ‘k’ is replaced by ‘s’ (2) ‘e’ is replaced by ‘i’ and (3) ‘g’ is added to the end. In calculating the distance between two tweets, we normalize their distance value by dividing it by the string length of the longer tweet. If tweets in the same category tend to use the same keywords, within-category SED should become smaller than between-category SED. We used Bibiko’s R package [22] for calculating the between-tweet distances.

2.3 Latent Semantic Analysis

The Latent Semantic Analysis (LSA) method is a broadly applied text processing technique [20]. LSA represents a set of tweets in a term by tweet matrix. A row in the matrix is a unique term. A column in the matrix is a tweet. Each matrix cell contains the frequency of each term within each tweet. This term-by-tweet matrix goes through singular value decomposition. Like principal component analysis, the factors are ordered by the amount of variance they capture in the original matrix. By using only the most influential factors, one can create an approximation of the original matrix, which removes the noise associated with the particular text sample, and uncovers somewhat abstract commonalities in word usage patterns. The approximated matrix yields a vector representation of terms with dimensionality equal to the number of factors included. The pair-wise similarities of all tweets are calculated by taking the vector cosine of the two tweets’ vectors.

Essentially, LSA exposes the similarity relations among related words by measuring how often these words appear together. In doing so, it reduces the dimensionality from thousands (i.e., the number of unique words in all documents) to hundreds. Tweets, however, are much shorter than typical documents used in LSA.

We used the LSA package for R by Wild [23]. We did not use the stemming option that reduced the words to the word-stems; we did use the English stop-words list provided in the package. An approximated term-by-tweet matrix was obtained. Because our analyses included 100 tweets or less, we used the few dimensions whose sum of singular values equaled or exceeded half of sum of singular values of all dimensions. Cosine similarities, which ranged between -1 and 1, were transformed to distances by subtracting these similarities from one and adding an epsilon value.

3 Method and Results

As an initial test, we applied the SED, LSA, and Wikipedia models to two sets of tweets that can be easily classified into three categories by humans. The first set had 45 tweets, consisting of 15 tweets from each of the three events that occurred at the time of our data collection: (1) death of J. D. Salinger, an American author, (2) an earthquake in Haiti, and (3) the release of iPad, Apple’s tablet computer. In the first set, the categories of all tweets were known.

The second set included all the tweets from the first set, and an additional 55 tweets that were randomly selected from the same time period. The addition of randomly sampled tweets tested the robustness of the classification techniques in a noisier environment.

The tweets were pre-processed by replacing any non-alphanumeric characters with the space character. The tweet-by-tweet distance matrix was obtained for each technique as described previously. A two-dimensional solution from multidimensional scaling of the distance matrix was used to predict the category membership of tweets in discriminant function analysis. We compared the techniques by looking at their accuracies of classifying tweets based on true positives and true negatives, using leave-one-out cross validation.

Table 1. Accuracy (hit plus correct rejection) of classifying 45 tweets with known categories

Technique	J. D. Salinger	iPad	Haiti
String Edit Distance	.67	.13	.60
Latent Semantic Analysis	.67	.73	.80
Wikipedia	.93	.87	.80

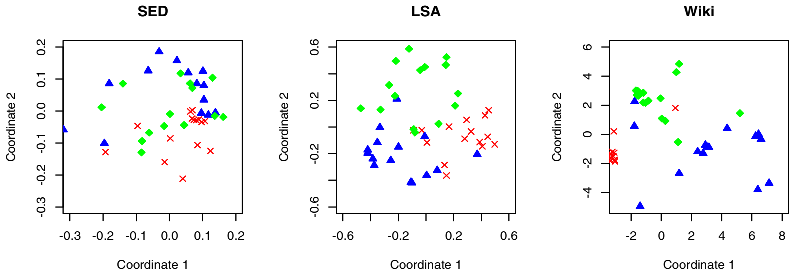

Fig. 4. Forty-five tweets with known categories mapped onto two-dimensional planes using multidimensional scaling of the between-tweet distances based on String Edit Distance, LSA and Wikipedia. An *x* is a tweet about J. D. Salinger and a triangle is a tweet about the iPad.

3.1 Tweets with Known Categories

Table 1 summarizes the classification accuracy of each technique. LSA and the Wikipedia model clearly performed better than SED. As shown in the Fig. 4, the Wikipedia distance measure yielded better-delineated clusters of related tweets than LSA. Interestingly, the Wikipedia model yielded a tight cluster for the tweets about J. D. Salinger, which generally discussed the unique topic of the author' s death; on the other hand, tweets about the iPad addressed a looser assortment of topics and, thus, clustered more loosely. In the Wikipedia distance space, there was one J. D. Salinger tweet that was far apart from other J. D. Salinger tweets. This tweet contained neither the author's name nor the title of his best-known book, which other tweets mentioned.

3.2 Adding Randomly Sampled Tweets

Table 2 summarizes the performances of LSA and Wikipedia for the data set containing randomly selected tweets. SED was dropped because of its weak performance in the first data set. When randomly sampled tweets were added, the Wikipedia model clearly outperformed LSA.

LSA's performance significantly deteriorated; that is, LSA was very sensitive to the addition of the other tweets. LSA processes terms in relation to what other terms appear in the corpus; thus LSA is highly affected by the context. On the other hand, the Wikipedia distance measure is robust to the addition of the randomly selected tweets, because the distance calculation is based on the Wikipedia page that best matches the topic of a given tweet, and the matching pages for existing tweets will not be affected by the introduction of new tweets. As can be seen in Fig. 5, the categories derived from the Wikipedia distance show clear separation, in contrast to the categories derived from LSA.

There was one randomly sampled tweet that was close to the tweets about J. D. Salinger. We thought this would be another tweet about the author, but it was not. This randomly sampled tweet was about the football player, "Warner." This tweet resulted in a short Wikipedia distance to tweets about J. D. Salinger, because the author is associated with Warner Books and Warner Brothers.

Table 2. Accuracy (hit plus correct rejection) of classifying 45 tweets with known categories when 55 randomly sampled tweets are added

Technique	J. D. Salinger	iPad	Haiti
Latent Semantic Analysis	.60	.60	.20
Wikipedia	.93	.87	.73

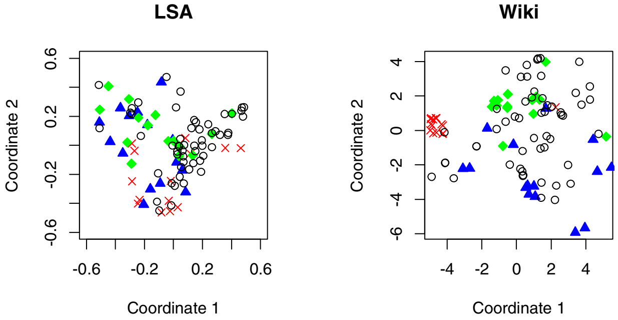

Fig. 5. Forty-five Tweets with known categories and 55 randomly selected tweets mapped onto two-dimensional planes using multidimensional scaling of the between-tweet distances based on LSA and Wikipedia. An open circle is a randomly sampled tweet, an *x* is a tweet about J. D. Salinger, and a triangle is a tweet about the iPad.

4 Discussion and Future Directions

A unique aspect of our technique is that it uses Wikipedia as its knowledge base to calculate between-tweet distances. Tweets and Wikipedia pages are both socially constructed artifacts. This, we think, allowed our technique to simulate the way humans categorize.

For our system to be used in production, the classification technique will need to be able to classify texts into events in near-real time. This, we think, is achievable: Once the novel tweet is mapped into the semantic space, the classifier can predict its category membership based on its similarity to other transformed tweets in the space.

Also, the classification technique will need to be adaptive. Although the randomly sampled tweets were treated as noise in the current work, seemingly useless tweets may actually contain useful information depending on the context. Future work might weight the value of information based on what topics are being discussed. Such classification models with selective attention mechanisms have been successful in simulating human classification behavior [24, 25].

In addition, the method for calculating distances using Wikipedia can be improved. The distances were purely based on the number of steps from one Wikipedia category to the other. The number of Wikipedia categories in each level, which was ignored in the current work, could be used to normalize the distances. In addition, LSA could be used to provide another distance measure between the Wikipedia categories, thus combining LSA and Wikipedia. The integration of probabilistic topic modeling techniques (e.g., [26]) might also be considered.

To recapitulate, when monitoring world events, the volume of tweets presents us with a problem: there is too much information to pay attention to. We are interested in looking at only the novel and useful information. In order to do so, we need ways of flagging emerging topics of interest, without a priori knowing what the topics will be. We suggest here an approach: short tweets are used to find longer passages in Wikipedia. These longer passages have already been linked to other Wikipedia passages. Thus, the distance between tweets can be approximated by the link distance measure between their corresponding Wikipedia pages. In an exploratory study, we showed this technique produced better classification accuracy than two other techniques, String Edit Distance and Latent Semantic Analysis. This work is an instance of a broader approach: by tapping Wikipedia and other living artifacts of social computing, computational methods might provide results that better serve humans.

Acknowledgments. We thank Valentin Leon-Bonnet and colleagues in the Center for Decision Technologies for their ideas.

References

1. McNamara, D.S.: Computational Methods to Extract Meaning From Text and Advance Theories of Human Cognition. Topics in Cognitive Science 3(1), 3–27 (2011)
2. Twitter blog. (2010), `http://blog.twitter.com/2010/02/measuring-tweets.html`
3. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: News in tweets. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 42–51. ACM, New York (2009)
4. Demirbas, M., Bayir, M.A., Akcora, C.G., Yilmaz, Y.: Crowd-sourced Sensing and Collaboration Using Twitter. In: 11th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), IEEE Computer Society Press, Los Alamitos (2010)
5. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on the World Wide Web, pp. 851–860. ACM, New York (2010)
6. Weick, K.E.: Sensemaking in organizations. Sage Publications, Inc., Thousand Oaks (1995)
7. Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: A first look. In: Proceedings of the Workshop on Analytics for Noisy, Unstructured Text Data (2010)

8. Macskassy, S.A.: Leveraging contextual information to explore posting and linking behaviors of bloggers. In: International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 64–71. IEEE, Los Alamitos (2010)
9. Gabrilovich, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research 34, 443–498 (2009)
10. Bratus, S., Rumshisky, A., Magar, R., Thompson, P.: Using domain knowledge for ontology-guided entity extraction from noisy, unstructured text data. In: Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data, pp. 101–106. ACM, New York (2009)
11. Strube, M., Ponzetto, S.P.: WikiRelate! Computing semantic relatedness using Wikipedia. In: Proceedings of the National Conference on Artificial Intelligence, p. 1419. AAAI Press, MIT Press (2006)
12. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (2010)
13. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceeding of the 17th International Conference on World Wide Web, pp. 91–100. ACM, New York (2008)
14. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009)
15. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 841–842. ACM, New York (2010)
16. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology 60(11), 2169–2188 (2009)
17. Stone, B., Dennis, S., Kwantes, P.J.: Comparing Methods for Single Paragraph Similarity Analysis. Wiley Online Library, Chichester (2010)
18. Venables, W.N., Ripley, B.D.: Modern applied statistics with S. Springer, Heidelberg (2002)
19. Ristad, E.S., Yianilos, P.N.: Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(5), 522–532 (2002)
20. Dumais, S.T., Landauer, T.K.: A solution to Platos problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological review 104, 211–240 (1997)
21. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on Twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining, pp. 1–10. ACM, New York (2010)
22. Bibiko, H.-J.: R code for Levensthein distance (2006)
23. Wild, F.: Latent Semantic Analysis Package in R (2010)
24. Kruschke, J.K.: ALCOVE: An exemplar-based connectionist model ot category learning. Connectionist psychology: a text with readings 99(1), 107 (1999)
25. Love, B.C., Medin, D.L., Gureckis, T.M.: SUSTAIN: A network model of category learning. Psychological Review 111(2), 309–332 (2004)
26. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(1), 5228 (2004)