Improving classification of tweets using word-word co-occurrence information from a large external corpus

Hugo Lewi Hammer
Department of Computer Science
Oslo and Akershus University College of Applied Sciences
Norway
hugo.hammer@hioa.no

Anis yazidi
Department of Computer Science
Oslo and Akershus University College of Applied Sciences
Norway
anis.yazidi@hioa.no

Aleksander Bai
Department of Computer Science
Oslo and Akershus University College of Applied Sciences
Norway
aleksander.bai@hioa.no

Paal Engelstad
Department of Computer Science
Oslo and Akershus University College of Applied Sciences
Norway
paal.engelstad@hioa.no

ABSTRACT

Classifying tweets is an intrinsically hard task as tweets are short messages which makes traditional bags of words based approach inefficient. In fact, bags of words approaches ignores relationships between important terms that do not co-occur literally.

In this paper we resort to word-word co-occurence information from a large corpus to expand the vocabulary of another corpus consisting of tweets. Our results show that we are able to reduce the number of erroneous classifications by 14% using co-occurence information.

CCS Concepts

•Information systems → Data mining; Web searching and information discovery; Social networks; •Applied computing → *Document management and text processing;*

Keywords

classification; lasso regression; twitter; word-word co-occurence

1. INTRODUCTION

SAC 2016,April 04-08, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2851613.2851986

Founded in 2006, Twitter (`www.twitter.com`) has grown to become one of the most popular social media services, known for its 140-character restriction on each post. In addition to a large general user base, Twitter is used extensively by celebrities, politicians, and news services to entertain, engage, or inform their followers. With over 500 million users, Twitter sees a daily stream of more than 400 million tweets a day [19].

Twitter is known to be an important source for early detecting of important events like breaking news, changes in the stock market, spread of diseases, earthquakes etc or analyzing different trends in politics, fashion, entertainment etc, see e.g. [14, 18, 9, 10, 17]. Such approaches are typically based on training a machine learner on a bag-of-words representation of the tweets, maybe in addition to other features like number of words, publication time etc. The bag of words representation is often unsatisfactory as it ignores relationships between important terms that do not co-occur literally. Many important words and phrases for correct classification may never occur in the training material, but only show up in the test material (e.g. future tweets). A bag-of-words approach will not be able to detect such tweets since the important words never occurred in the training set. For example, suppose we want to detect tweets about the war in Syria. In the manually annotated training material we may have good predictors like "al-Assad", "Syria", "Homs" etc, but may miss other relevant phrases like "Damascus", "gas attack", "Baath party", "ISIL" which potentially could improve the classifications since such words are likely to occur in future tweets about the Syrian war.

In this paper we suggest to "enrich" the vocabulary in the training material with other potentially relevant phrases by using word-word co-occurrence information from an other large news corpus (1.1 billion words). Computing words that tend to co-occur with "al-assad" in the news corpus,

we find among the top ten words "bashar", "al-sharaa" (vice president in Syria), "negotiations" and "syria" which seem like other relevant words to detect tweets about the Syrian war. It's not obvious what's the best way to incorporate such external co-occurrence information in the training material of tweets. We therefore suggest a large set of different approaches and test them extensively on real twitter data.

2. RELATED WORK

Techniques for enriching text fall under two main categories: those who use intrinsic information contained in the current corpus and those who use extern resources. A representative example of intrinsic techniques is the the Self-Term Expansion Methodology due to Pinto et al. [15] for clustering tweets. The method compromises two main steps: the Self-Term Enriching step, and a Term Selection step. The Self-Term Enriching procedure enriches the text representation of the tweets by exploiting the current tweets corpus and without the need of any external corpus, that is why the technique is called Self-Term Enriching. Terms of a documents are represented with a set of co-related terms. A co-occurrence list is calculated from the target data set by applying Pointwise Mutual Information (PMI). The Term Selection step identifies the most important features and tries to reduce the noise introduced by the the Self-Term Enriching phase.

The second category of techniques for enriching text representation uses external resources other than the current text materials to be clustered or classified. It is worth mentioning that the later techniques have received most attention in the literature compared to techniques that resort to intrinstic information for the enriching task. For example in [7, 8, 16, 4], the authors enrich the text representation using WordNet [12] where terms of the documents are replaces with their hypernym and synonym.

Similarly, the seminal work of Gabrilovich et al. [6] leverages knowledge bases from Wikipedia and Open Directory Project (ODP) in order to enhance the textual representation of short messages. The authors concluded that augmented knowledge based features generated from ODP and Wikipedia improved the text categorization task.

Alahmadi et al. [1] use an approach based on supplementing the bag-of-words representational scheme with a concept-based representation that utilises Wikipedia as a knowledge base.

In [2] Wikipedia semantic knowledge are used to tackle data sparseness in a question answering task. Experiments show that the approach significantly outperforms the baseline method (with error reductions of 23.21%).

Chen et al. [3] propose a word-word co-occurrence matrix based method for improved relevance feedback in information retrieval. Unlike other studies about word association, the authors consider the influence of the inter word distance and co-windows ratio. Experiments with TREC dataset demonstrate the effectiveness of the method.

3. WORD-WORD CO-OCCURRENCE MATRIX AND DOCUMENT TERM MATRIX

In this section we represent relevant background information for the rest of the paper. More specifically we define the word-word co-occurrence matrix (COM) and the document term matrices (DTM).

3.1 Word-word co-occurrence matrix

Suppose we have a large corpus consisting of a total of N words and let $w_1, w_2, \ldots, w_{N_w}$ denote the different unique words in the corpus. Further let $N_i,\ i \in \{1, 2, \ldots, N_w\}$ denote the number of times w_i occurs in the corpus and let $N_{ij},\ i, j \in \{1, 2, \ldots, N_w\}$ denote the number of times w_i occurs in the neighbourhood of w_j in the corpus. The neighbourhood of a word, w_j, is typically those words closest to w_j in front and behind in the text. We assume symmetry such that $N_{ji} = N_{ij}$. A COM is the matrix with the element N_{ij} in position $(i, j),\ i, j = 1, 2, \ldots, N_w$. A COM computed from a large corpus is a highly valuable tool to analyze semantic relations between words, see e.g. [11, 13].

Suppose we want to use COM to compute the semantic relation between w_i and w_j. There are typically three main approaches: Correlation, angle and PMI between words in COM. In our experiments the PMI performed better than the other two approaches and the descriptions below therefore are based on PMI.

3.2 Document term matrix

Other words for a document term matrix (DTM) are bag-of-words and n-grams. Suppose that a corpus consist of D tweets (more generally documents). Let n_{di} denote the number of times word w_i occur in tweet $d \in \{1, 2, \ldots, D\}$ and n_w the total number of unique words in the D tweets. A DTM is the matrix with the elements n_{di} in positions $(d, i),\ d = 1, 2, \ldots, D, i = 1, 2, \ldots, n_w$. A natural generalization is to not only use words, but all phrases of subsequent words in the corpus called n-grams. In this paper we only resort to single words (unigram). Reweightings of the pure term frequencies in a DTM is also very common, e.g. the TF-IDF ([11], chapter 15).

4. INCORPORATING CO-OCCURRENCE INFORMATION

In this section we present a framework to incorporate COM information from a large external corpus to a DTM. We start by expanding the vocabulary of DTM from all the unique words in the tweets to the union of the unique words in the tweets and the words in COM. See Figure 1 for a simple visualization of the expansion. The gray part shows the additional words added to the original DTM shown as the white part of the matrix. Our goal is to add reasonable values in the gray part of the matrix and adjust values in the white part of the matrix to improve classification. To simplify the notation below, let r_{ij} refer to $\widehat{PMI}(w_i, w_j)$. Also assume that all words in the tweet vocabulary are part of the COM vocabulary. In practice we obtained this by letting words that is in the tweet vocabulary and not in the COM vocabulary, are added to COM with all co-occurence

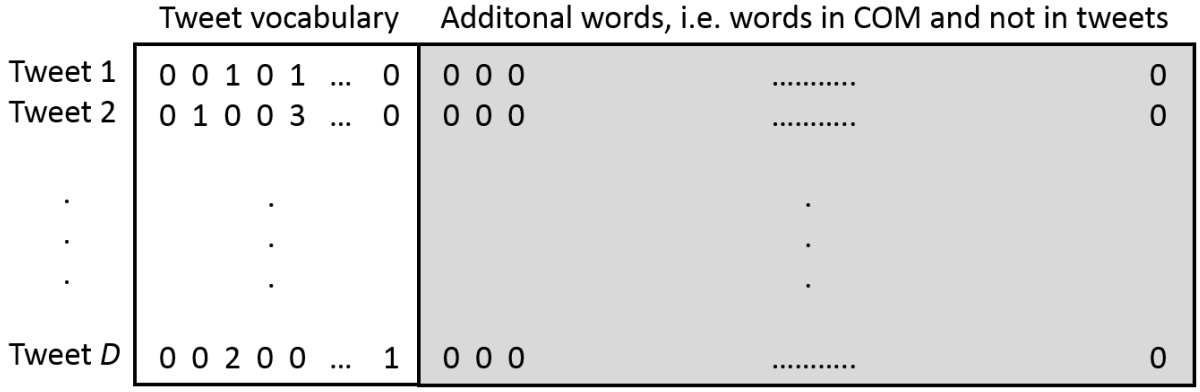

Figure 1: Illustration of the expansion (shown in gray) of the original tweet DTM shown in white.

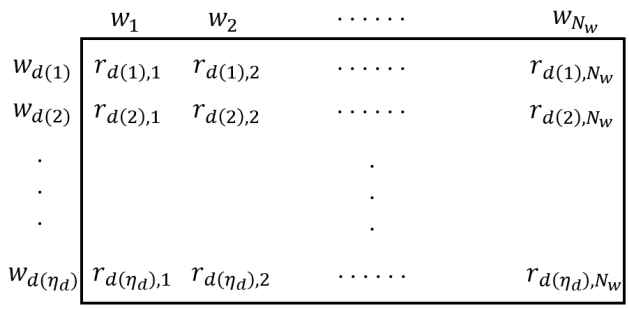

Figure 2: Illustration of the of the matrix $\mathbf{PMI_{tweet}}$.

frequencies with other words equal to zero.

Suppose a tweet $d \in \{1, 2, \ldots, D\}$ consists of the η_d unique words $w_{d(1)}, \ldots, w_{d(\eta_d)}$ and recall that we assume that all being part of the COM vocabulary $w_1, w_2, \ldots, w_{N_w}$. Further let $n_{d,d(1)}, \ldots, n_{d,d(\eta_d)}$ denote the frequency (or some reweighting like TF-IDF) of $w_{d(1)}, \ldots, w_{d(\eta_d)}$. Define the matrix $\text{PMI}_{\text{tweet}}$ consisting of the entries $r_{d(i),j}$, $i = 1, 2, \ldots, \eta_d$, $j = 1, 2, \ldots, N_w$ containing the PMI scores between the words in the tweet and all the words in COM. Figure 2 illustrates this matrix. Based on $\text{PMI}_{\text{tweet}}$ we can expand the vocabulary of the tweet d in different ways. Maybe the most natural is for each word in COM to compute the sum of PMI scores for the words in the tweet and add this values to the expanded DTM shown in Figure 1

$$\widetilde{n}_{d,j} = \frac{1}{\eta_d} \sum_{i=1}^{\eta_d} n_{d,d(i)} (r_{d(i),j})^{\gamma} \qquad (1)$$

The parameter γ control if the sum of many fairly high values of $r_{d(i),j}$ result in a higher score than one very high value.

5. LINGUISTIC RESOURCES

The COM are computed from a huge corpus that is made openly available by the National Library of Norway (NLN). The corpus consists of news articles collected from Norwegian newspapers from 1998 until 2011. This corresponds to roughly 1.1 billion Norwegian words distributed over 4 million articles. To compute N_{ij}, we used a neighborhood of six words in front and behind of w_j (recall Section 3.1). We only used words that occurred at least 50 times in the news corpus ending up with a vocabulary with 287904 unique words.

The Twitter corpus is selected from all tweets published in Norwegian on Twitter from 20th of July to 8th of August 2011 a total of about two million tweets. We selected a subset of tweets as follows:

1. We counted the number of times different hashtags were used.
2. Among the most frequently used hashtags we manually picked hashtags related to six topics and selected all the tweet consisting at least one of these hashtags. Examples of hashtags were #Utøya and #PrayForOslo related to the 22th July 2011 terror, #Libya and #Bieber.

The resulting corpus consists of a total of 21270 tweets. The classification task is to classify the correct topic of these tweets when all the hashtags are removed from the tweets.

6. CLASSIFICATION PERFORMANCE

In this section we compare the classification performance of using (1) compared to using the original DTM. We base our classifications on multinomial LASSO regression [5]. We expect that incorporating external information is particularly useful if the number of documents (tweets) in the annotated training material are few. Then many important predictors (words) are missing in the training material and thus not being part of the classifier. Our results is in accordance with this. Using 30% or more of the tweet corpus to train the classifier (more than 6381 tweets), the reduction in erroneous classifications is below 5% compared to not using external information. Using less than 30% of the tweets to train the classifier, the reduction in erroneous classifications is between 5 and 15%. The best results were achieved setting $\gamma = 0.1$ or $\gamma = 10$ in (1) compared to setting $\gamma = 1$

As expected a higher percentage of the tweets are classified correctly when 10% of the tweets are used for training com-

pared to only 5%. For 10% training the highest reduction in erroneous classifications were

$$\frac{(100-73.1)-(100-75.9)}{100-73.1}\cdot 100\% = 10.4\%$$

For 5% training the highest reduction is

$$\frac{(100-69.0)-(100-73.4)}{100-69.0}\cdot 100\% = 14.2\%$$

We see, as expected, that when the training set is small inclusion of external co-occurence information have a larger positive effect on the classification performance. An other interesting observation is that using 10% training with no external information performs poorer (73.1%) than using 5% training and external information (73.4%). In other words it is better to include external co-occurence information than increasing the number of annotated tweets from 5% (1064 tweets) to 10% (2127 tweets). Having in mind that manual annotation of documents are very resource demanding, this is quite an impressive result and documents the usefulness of the method in this paper.

7. CLOSING REMARKS

In this paper we have shown how external information from a word-word co-occurrence matrix can be used to improve the classification of tweets.

8. REFERENCES

[1] A. Alahmadi, A. Joorabchi, and A. Mahdi. A new text representation scheme combining bag-of-words and bag-of-concepts approaches for automatic text classification. In *GCC Conference and Exhibition (GCC), 2013 7th IEEE*. IEEE Press, 2013.

[2] L. Cai, G. Zhou, K. Liu, and J. Zhao. Large-scale question classification in cqa by leveraging wikipedia semantic knowledge. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1321–1330, New York, NY, USA, 2011. ACM.

[3] Z. Chen and Y. Lu. A word co-occurrence matrix based method for relevance feedback. *Journal of Computational Information Systems*, 7(1):17 – 24, 2011.

[4] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[6] E. Gabrilovich and S. Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *AAAI*, volume 6, pages 1301–1306, 2006.

[7] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 541–544. IEEE, 2003.

[8] A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In *Proc. SIGIR Semantic Web Workshop*, 2003.

[9] V. Lampos, T. De Bie, and N. Cristianini. Flu detector: Tracking epidemics on twitter. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, pages 599–602, Berlin, Heidelberg, 2010. Springer-Verlag.

[10] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter trending topic classification. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pages 251–258, Washington, DC, USA, 2011. IEEE Computer Society.

[11] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[12] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[13] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors forword representation. http://nlp.stanford.edu/projects/glove/glove.pdf, 2015. [Online; accessed 27-July-2015].

[14] S. Petrović, M. Osborne, and V. Lavrenko. Using paraphrases for improving first story detection in news and twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 338–346, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[15] D. Pinto, P. Rosso, and H. Jiménez-Salazar. A self-enriching methodology for clustering narrow domain short texts. *The Computer Journal*, 54(7):1148–1165, 2011.

[16] M. Rodriguez, J. Hidalgo, and B. Agudo. Using wordnet to complement training information in text categorization. In *Proceedings of 2nd International Conference on Recent Advances in Natural Language Processing II: Selected Papers from RANLP*, volume 97, pages 353–364, 2000.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[18] X. Zhang, H. Fuehres, and P. A. Gloor. Predicting Stock Market Indicators Through Twitter I hope it is not as bad as I fear. *Procedia - Social and Behavioral Sciences*, 26:55–62, Jan. 2011.

[19] A. Zubiaga, D. Spina, R. Martinez, and V. Fresno. Real-time classification of twitter trends. *Journal of the American Society for Information Science and Technology*, 66(3):462 – 473, 2015.