

Discovering Health-Related Knowledge in Social Media Using Ensembles of Heterogeneous Features

Suppawong Tuarob

Conrad S Tucker

Marcel Salathe

Nilam Ram

The Pennsylvania State University

To appear in CIKM 2013. San Francisco, CA

Oct 27th – Nov 1st 2013


Motivation

- Misclassification of the state-of-the-art method using uni, bi, tri grams with SVM classifier.
 - Keyword Recognition Problem.
 - `yep he's fine...was only a mild case of the swine :)`
 - Term Disambiguation Problem.
 - This is `sick` , it's snowing again. :- It's like i am living in Russia.
- Traditional document classification techniques would fail when dealing with social media because:
 - They are high-dimensional but sparse: due to having short length.
 - They are noisy: Grammatical errors, misspelling, new terms.




Research Objectives

- A message is said to be ***health-related*** if at least one of these two following conditions is met:
 - The message indicates its author has health issues.
 - The message talks about someone else getting sick, or expresses health concern.



Fever, back pain,
headache... ugh!



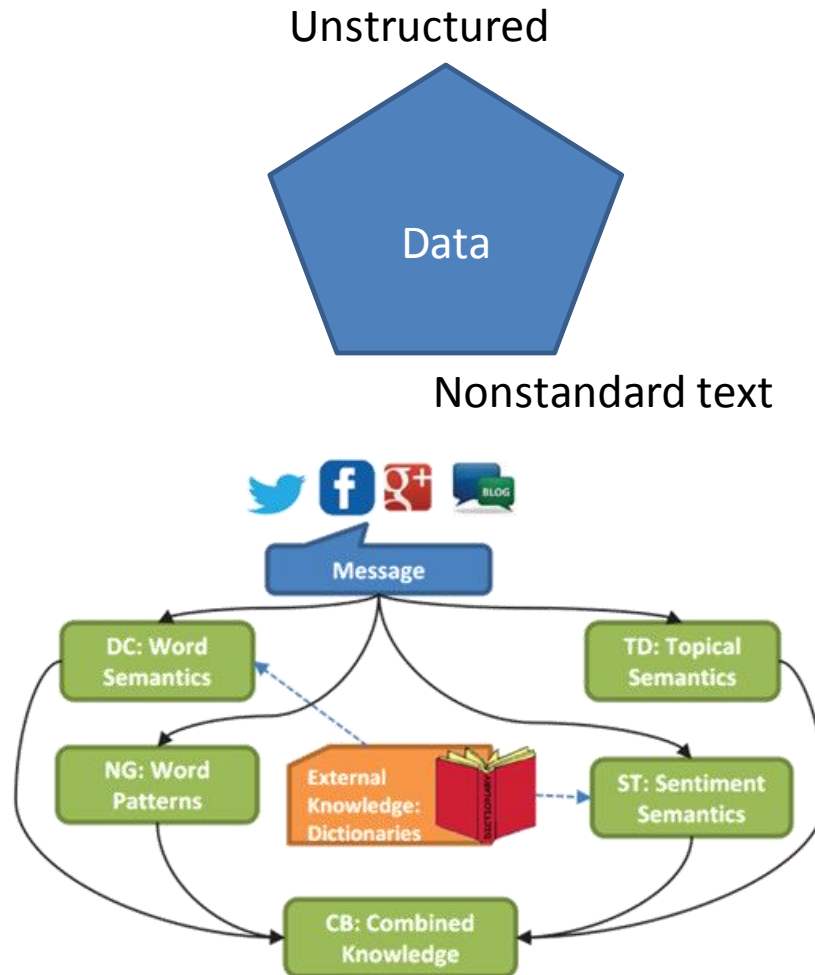
I completely
understand, more
than anyone! Try a
warm bath too.
That always helped
me w/ Pauly. &
drinking water.

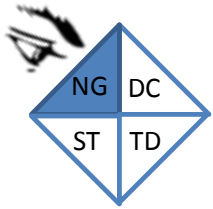
Previous Works on Social Media Document Classification

- Keyword Based
 - Ginsberg et al., Culotta, Corley et al. identified flu-related content in query logs [17,28,29].
 - Yang et al. identified content containing the adverse drug reactions [24,30].
- Learning Based
 - N-gram based classification [21,22,32] (Baseline)
 - Keyword filtering -> N-gram based classification [5]
 - Social media specific features: Authors and reply-to users [33]

Methodology: Overview

- 5 different feature types representing semantically different aspect of the data.
- A machine is trained to learn a different aspect.
- Combine 5 base classifiers using standard ensemble methods.





N-Gram Features (NG)

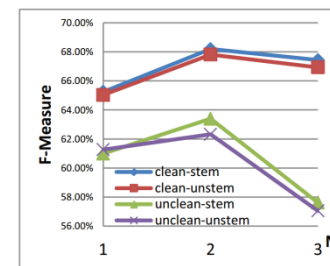
- Represent a document with N-grams.
- N-gram features have been used extensively in text classification to learn word patterns in the training data.
- Best configuration:
 - $\langle c = \text{SVM}, \text{clean} = \text{T}; \text{stem} = \text{T}; N = 2; W = \text{tfidf} \rangle$
- Baseline by Paul and Dredze [21]:
 - $\langle c = \text{SVM}, \text{clean} = \text{F}; \text{stem} = \text{F}; N = 3; W = \text{binary} \rangle$

Param.	Description	Possible Values
clean	whether to remove punctuation and lowercase the message	T,F
stem	whether to apply Porter's stemming algorithm to the message	T,F
N	Max number of consecutive terms to form grams	1,2,3
W	Weighting schemes	binary, freq, tfidf

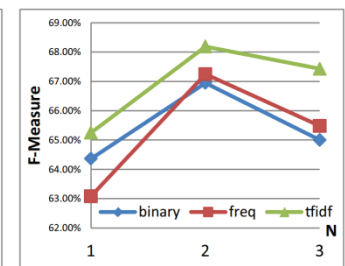
$$f_i^{\text{bin}} = \begin{cases} 1 & ; \text{if } v_i \in t \text{ and } v_i \in V \\ 0 & ; \text{otherwise} \end{cases}$$

$$f_i^{\text{freq}} = TF(v_i, t)$$

$$f_i^{\text{tfidf}} = \begin{cases} \frac{TF(v_i, t)}{\max(TF(w, t); w \in t)} \cdot \log \frac{|S|}{1 + |s \in S; v_i \in s|} & ; \text{if } v_i \in t \\ 0 & ; \text{otherwise} \end{cases}$$

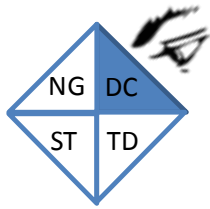


(a) Varying clean/stem parameters



(b) Varying weight (W) parameters

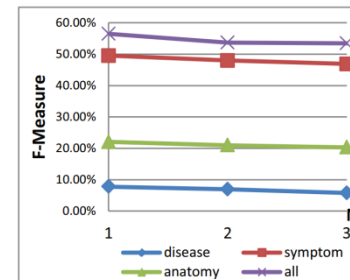
Parameter comparison of NG feature extraction as the maximum size of grams (N).



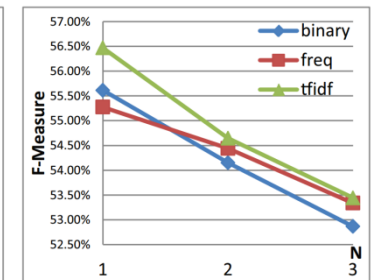
Dictionary Based Compound Features (DC)

- Problems with NG features:
 - Words with multiple meaning are treated the same (Ex. **cold** can be used in both disease or temperature contexts)
 - Important keywords are treated as normal words (Ex. **Xeroderma pigmentosum**)
- Represent a document with *compounds* [47], each of which must contain at least a keyword from the dictionary.
- Best configuration (F = 56.47%):
 - $\langle c = \text{SVM}, \text{stem} = \text{true}; \text{vocab} = \text{all}; N = 1; C = 2; W = \text{tfidf} \rangle$

Param.	Description	Possible Values
stem	whether to apply Porter's stemming algorithm to the message	T,F
vocab	Vocabularies used	disease, symptom, anatomy, all
N	Max number of consecutive terms to form grams	1,2,3
C	Maximum number of terms in a compound	1,2
W	Weighting schemes	binary, freq, tfidf

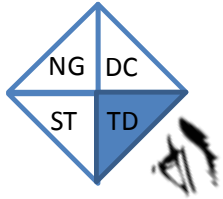


(a) Varying vocab parameters



(b) Varying weight (W) parameters

Parameter comparison of DC feature extraction as the function of maximum gram size (N).



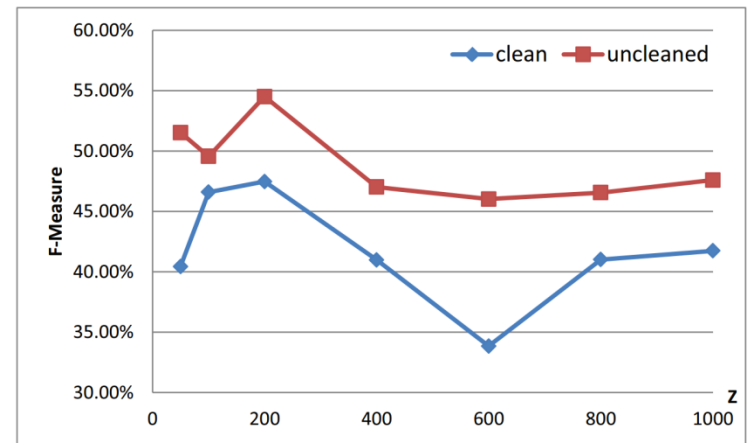
Topic Distribution Features (TD)

- Represent a document with topic distribution.
- Use LDA to model topics.

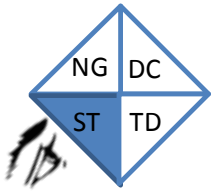
$$P(w_i|d) = \sum_{j=1}^{|Z|} P(w_i|z_i = j) \cdot P(z_i = j|d)$$

- Best configuration:
 - $\langle c = \text{Random Forest}; \text{clean} = \text{F}; Z = 200 \rangle$

Param.	Description	Possible Values
clean	Whether to remove punctuation and stopwords, stem the message	T,F
Z	Number of topics	50, 100, 200, 400, 600, 800, 1000



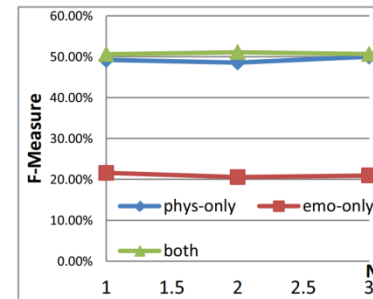
Parameter comparison of TD feature extraction as the function of number of topics (Z)



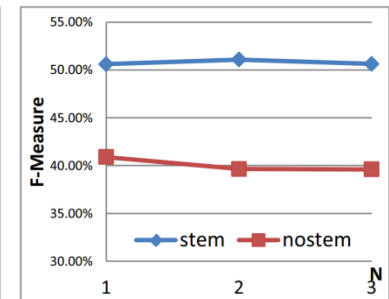
Sentiment Features (ST)

- Physical Based:
 - Number of health-related keywords
- Emotion Based:
 - Positive/Negative sentiment scores
- Best Configuration (F = 51.08%):
 - $\langle c = \text{RIPPER}; \text{stem} = \text{T}; N = 2; \text{type} = \text{both} \rangle$

Grp	Feature Name	Description
Phys	num disease-words	Number of disease terms
	ratio num diseasewords	Ratio of – to all terms
	num symptomwords	Number of symptom terms
	ratio num symptomwords	Ratio of – to all terms
	num anatomywords	Number of anatomy terms
	ratio num anatomywords	Ratio of – to all terms
	num healthwords	Number of health-related words
Emo.	ratio num healthwords	Ratio of – to all terms
	positive emotion	Positive Emotional Level (1-5)
	negative emotion	Negative Emotion Level (1-5)
	num pos emoticons	Num positive emoticons, e.g. :), (:]
	num neg emoticons	Num negative emoticons, e.g. :(, =(



(a) Varying type parameter



(b) Varying stem parameter

Parameter comparison of ST feature extraction as the function of maximum gram size (N)

Base Classifiers

- **Random Forest (RF)**[38] is a tree-based ensemble classifier consisting of many decision trees.
- **Support Vector Machine (SVM)**[40] is a function based classifier built upon the concept of decision planes that define decision boundaries.
- **Repeated Incremental Pruning to Produce Error Reduction (RIPPER)**[42] is a rule-based classifier which implements a propositional rule learner.
- **NaiveBayes (NB)**[43] is a simple probabilistic classifier implementing Bayes' theorem. NaiveBayes has been shown to perform superior in some text classification tasks such as spam filtering [44].

Ensemble Methods

- **Majority Voting (VOTE)** Each classifier outputs either a 'yes' or 'no'. The final outcome is the majority vote of all the classifiers.
- **Weighted Probability Averaging (WPA)** Each classifier is given a weight, where the sum of all weights is 1. Each classifier outputs a probability estimate of the positive class. The final output is the weighted average of all the classifiers.
- **Multi Staging (MS)** Classifiers operate in order. If a classifier says 'yes', the final output is yes; otherwise the instance is passed to the next classifier to decide.
- **Reverse Multi Staging (RevMS)** Similar to the MS technique, except that an instance is passed to the next classifier if the prior classifier says 'yes'.



Combined Features (CB)

- Having a classifier that learns all the aspects of the data may be helpful when combined with other one-aspect classifiers.
- We create such an overall classifier by training a base classifier with combined features generated by merging all the four feature sets discussed above into a single feature set with SVM as the base classifier.

Experiments and Classification Results

- Dataset: 5,128 manually labeled tweets
 - Positive: 1,832 (35.73%)
 - Negative 3,296 (64.27%)
- 10 Fold X-validation with 10% held-out data.

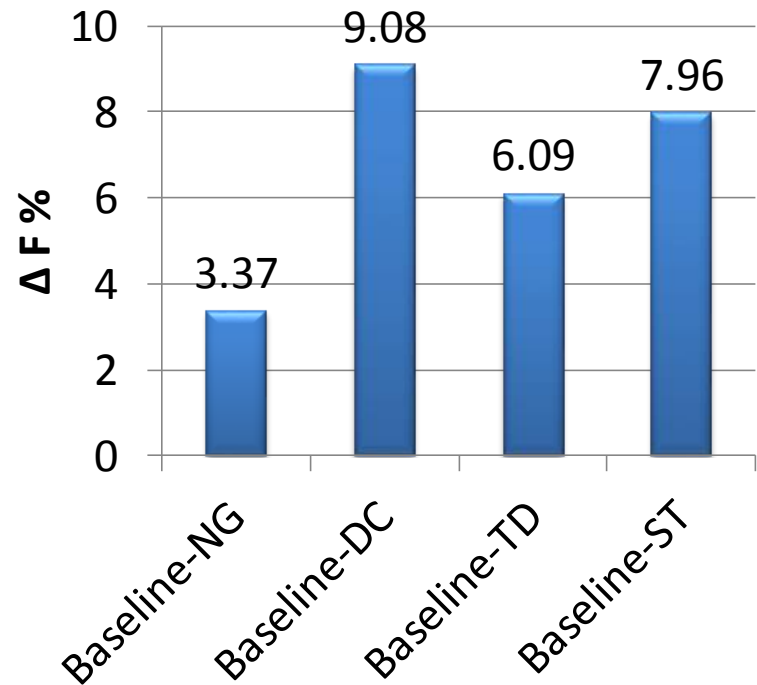
Classifier	Pr %	Re %	F1 %	$\Delta F1$ %
Baseline	76.68	47.63	58.76	0.00
NG	75.65	62.06	68.19	9.43
DC	73.77	45.74	56.47	-2.29
TD	70.48	44.43	54.50	-4.26
ST	55.87	47.05	51.08	-7.68
CB	85.07	57.29	68.47	9.71
VOTE	77.32	65.24	70.77	12.01
WPA	80.45	74.52	77.37	18.61
MS	56.51	91.93	69.99	11.23
RevMS	90.08	37.96	53.41	-5.35

10 fold classification performance of the baseline, proposed base and ensemble classifiers, in terms of precision, recall, F1, and $\Delta F1$ on the dataset

Impact of Each Feature Type

- Each proposed feature type is combined with the features used by the baseline.
- **NG features:** Impact is not significant since the baseline and our NG features are both N-gram based; hence, they provide redundant information to the classifier.
- **DC features:** Most impact on the performance, because it can mitigate both keyword-recognition and term-disambiguation problems.
- **ST features** capture both health-related keywords used and emotion in a document. Since these properties are not captured in the baseline feature set, combining the ST features with the baseline allows the classifier to learn more information as expected.

Performance impact of each proposed feature set on the baseline feature set



Contributions

- Develop a public health surveillance system using the dynamic large scale availability of social media data.
- Propose to use 5 heterogeneous feature types representing different aspects of semantics for identification of health-related messages in social media.
- Combine feature types using ensemble methods where each base classifier learns a different aspect of the data.

Conclusions

- Propose to use 5 semantically heterogeneous feature types for short text classification tasks.
- Propose to combine the features by combining base classifiers each of which learns a different aspect of the data using standard ensemble techniques.
- The proposed methodology outperforms the baseline using N-gram binary feature by 18.61%.
- Dictionary based compound features have the most additional impact since they can solve both keyword recognition and term disambiguation posed by the features used by the baseline.

References

- [1] C. Tucker, H. Kim, Predicting emerging product design trend by mining publicly available customer review data, Proceedings of the 18th International Conference on Engineering Design (ICED11) 6 (2011) 43–52.
- [2] S. Tuarob, C. S. Tucker, Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data, in: Proc. ASME 2013 Int. Design Engineering Technical Conf. Computers and Information in Engineering Conf., IDETC/CIE '13, 2013.
- [3] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: realtime event detection by social sensors, in: Proceedings of the 19th international conference on World wide web, WWW '10, 2010, pp. 851–860.
- [4] C. Caragea, N. McNeese, A. Jaiswal, G. Traylor, H. Kim, P. Mitra, D. Wu, A. Tapia, L. Giles, B. Jansen, et al., Classifying text messages for the haiti earthquake, in: Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM2011), 2011.
- [5] N. Collier, S. Doan, Syndromic classification of twitter messages, CoRRabs/1110.3094.
- [6] L. Lopes, J. Zamite, B. Tavares, F. Couto, F. Silva, M. Silva, Automated social network epidemic data collector, in: INForum informatics symposium. Lisboa, 2009.
- [7] P. Chira, L. Nugent, K. Miller, T. Park, S. Donahue, A. Soni, D. Nugent, C. Sandborg, Living profiles: Design of a health media platform for teens with special healthcare needs, Journal of biomedical informatics 43 (5) (2010) S9–S12.
- [8] P. F. Brennan, S. Downs, G. Casper, Project health design: Rethinking the power and potential of personal health records, Journal of Biomedical Informatics 43 (5, Supplement) (2010) S3–S5, ?ce:title?Project Health Design?/ce:title?. doi:http://dx.doi.org/10.1016/j.jbi.2010.09.001.
- [9] M. Merolli, K. Gray, F. Martin-Sanchez, Health outcomes and related effects of using social media in chronic disease management: A literature review and analysis of affordances, Journal of biomedical informatics.
- [10] M. Terry, Twittering healthcare: social media and medicine, Telemedicine and e-Health 15 (6) (2009) 507–510.
- [11] J. Kaye, L. Curren, N. Anderson, K. Edwards, S. M. Fullerton, N. Kanellopoulou, D. Lund, D. G. MacArthur, D. Mascalon, J. Shepherd, et al., From patients to partners: participant-centric initiatives in biomedical research, Nature Reviews Genetics 13 (5) (2012) 371–376.
- [12] B. Hesse, D. Hansen, T. Finholt, S. Munson, W. Kellogg, J. Thomas, Social participation in health 2.0, Computer 43 (11) (2010) 45–52. doi:10.1109/MC.2010.326.
- [13] S. H. Jain, Practicing medicine in the age of facebook, New England Journal of Medicine 361 (7) (2009) 649–651, PMID: 19675328. doi:10.1056/NEJMp0901277.
- [14] M. v. d. Eijk, J. M. Faber, W. J. Aarts, A. J. Kremer, M. Munneke, R. B. Bloem, Using online health communities to deliver patient-centered care to people with chronic conditions, J Med Internet Res 15 (6) (2013) e115. doi:10.2196/jmir.2476. URL http://www.jmir.org/2013/6/e115/
- [15] J. Greene, N. Choudhry, E. Kilabuk, W. Shrank, Online social networking by patients with diabetes: A qualitative evaluation of communication with facebook, Journal of General Internal Medicine 26 (3) (2011) 287–292. doi:10.1007/s11606-010-1526-3. URL http://dx.doi.org/10.1007/s11606-010-1526-3
- [16] A. Culotta, Detecting influenza outbreaks by analyzing twitter messages, CoRR abs/1007.4748.
- [17] C. Corley, D. Cook, A. Mikler, K. Singh, Using web and social media for influenza surveillance, in: H. R. Arabnia (Ed.), Advances in Computational Biology, Vol. 680 of Advances in Experimental Medicine and Biology, Springer New York, 2010, pp. 559–564.
- [18] T. Bodnar, M. Salath, e, Validating models for disease detection using twitter, in: Proceedings of the 22nd international conference on World Wide Web companion, WWW '13 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2013, pp. 699–702.
- [19] N. Heavilin, B. Gerbert, J. Page, J. Gibbs, Public health surveillance of dental pain via twitter, Journal of dental research 90 (9) (2011) 1047–1051.
- [20] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
- [21] M. J. Paul, M. Dredze, A model for mining public health topics from twitter, Tech. rep. (2011).
- [22] M. J. Paul, M. Dredze, You are what you tweet: Analyzing Twitter for public health, Fifth International AAAI Conference on Weblogs and Social Media (2011) 265–272.
- [23] D. Cameron, G. A. Smith, R. Daniulaityte, A. P. Sheth, D. Dave, L. Chen, G. Anand, R. Carlson, K. Z. Watkins, R. Falck, Predose: A semantic web platform for drug abuse epidemiology using social media, Journal of Biomedical Informatics (0) (2013) –. doi:http://dx.doi.org/10.1016/j.jbi.2013.07.007.
- [24] C. C. Yang, L. Jiang, M. Zhang, Social media mining for drug safety signal detection, in: Proceedings of the 2012 international workshop on Smart health and wellbeing, SHB '12, ACM, New York, NY, USA, 2012, pp. 33–40. doi:10.1145/2389707.2389714.
- [25] X.-H. Phan, L.-M. Nguyen, S. Horiguchi, Learning to classify short and sparse text & web with hidden topics from large-scale data collections, in: Proceedings of the 17th international conference on World Wide Web, WWW '08, 2008, pp. 91–100.
- [26] X. Hu, N. Sun, C. Zhang, T.-S. Chua, Exploiting internal and external semantics for the clustering of short texts using world knowledge, in: Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09, ACM, New York, NY, USA, 2009, pp. 919–928. doi:10.1145/1645953.1646071.
- [27] O. Jin, N. N. Liu, K. Zhao, Y. Yu, Q. Yang, Transferring topical knowledge from auxiliary long texts for short text clustering, in: Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11, ACM, New York, NY, USA, 2011, pp. 775–784. doi:10.1145/2063576.2063689.
- [28] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data., Nature 457 (7232) (2009) 1012–4. doi:10.1038/nature07634.
- [29] A. Culotta, Towards detecting influenza epidemics by analyzing twitter messages, in: Proceedings of the First Workshop on Social Media Analytics, SOMA '10, ACM, New York, NY, USA, 2010, pp. 115–122. doi:10.1145/1964858.1964874.
- [30] Q. T. Zeng, T. Tse, Exploring and developing consumer health vocabularies, Journal of the American Medical Informatics Association 13 (1) (2006) 24–29.
- [31] N. Collier, S. Doan, A. Kawazoe, R. Goodwin, M. Conway, Y. Tateno, Q. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, et al., Biocaster: detecting public health rumors with a web-based text mining system, Bioinformatics 24 (24) (2008) 2940–2941.

References

- [32] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the flu: detecting influenza epidemics using twitter, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 1568–1576.
- [33] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, M. Demirbas, Short text classification in twitter to improve information filtering, in: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, 2010, pp. 841–842.
- [34] K. Kira, L. Rendell, The feature selection problem: Traditional methods and a new algorithm, in: Proceedings of the National Conference on Artificial Intelligence, John Wiley & Sons Ltd, 1992, pp. 129–129.
- [35] A. Silvescu, C. Caragea, V. Honavar, Combining super-structuring and abstraction on sequence classification, in: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, ICDM '09, IEEE Computer Society, Washington, DC, USA, 2009, pp. 986–991. doi:10.1109/ICDM.2009.130.
- [36] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, Target-dependent twitter sentiment classification, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 151–160.
- [37] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment in short strength detection in informal text, J. Am. Soc. Inf. Sci. Technol. 61 (12) (2010) 2544–2558. doi:10.1002/asi.v61:12.
- [38] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.
- [39] T. M. Khoshgoftaar, M. Golawala, J. V. Hulse, An empirical study of learning from imbalanced data using random forest, in: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 02, ICTAI '07, 2007, pp. 310–317.
- [40] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [41] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, Springer, 1998.
- [42] W. W. Cohen, Fast effective rule induction, in: Twelfth International Conference on Machine Learning, Morgan Kaufmann, 1995, pp. 115–123.
- [43] G. H. John, P. Langley, Estimating continuous distributions in bayesian classifiers, in: Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo, 1995, pp. 338–345.
- [44] I. Androustopoulos, J. Koutsias, K. V. Chandrinos, C. D. Spyropoulos, An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages, in: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2000, pp. 160–167.
- [45] A. McCallum, K. Nigam, A comparison of event models for naive bayes text classification, in: AAAI-98 Workshop on 'Learning for Text Categorization', 1998.
- [46] A. McCallum, K. Nigam, et al., A comparison of event models for naive bayes text classification, in: AAAI-98 workshop on learning for text categorization, Vol. 752, Citeseer, 1998, pp. 41–48.
- [47] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Goncalves, W. M. Jr., Word co-occurrence features for text classification, Information Systems 36 (5) (2011) 843–858. doi:10.1016/j.is.2011.02.002.
- [48] S. Kataria, P. Mitra, S. Bhatia, Utilizing context in generative bayesian models for linked corpus, in: AAAI, 2010.
- [49] S. Tuarob, L. C. Pouchard, N. Noy, J. S. Horsburgh, G. Palanisamy, On mercury: Towards automatic annotation of environmental science metadata, in: Proceedings of the 2nd International Workshop on Linked Science 2012: Tackling Big Data, LISC '12, 2012.
- [50] S. Tuarob, L. C. Pouchard, C. L. Giles, Automatic tag recommendation for metadata annotation using probabilistic topic modeling, in: Proceedings of the 13th ACM/IEEE-CS joint conference on Digital Libraries, JCDL '13, 2013.
- [51] T. L. Griffiths, M. Steyvers, D. M. Blei, J. B. Tenenbaum, Integrating topics and syntax, Advances in neural information processing systems 17 (2005) 537–544.
- [52] D. D. Walker, W. B. Lund, E. K. Ringger, Evaluating models of latent document semantics in the presence of ocr errors, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10, 2010, pp. 240–250.
- [53] S. Kataria, P. Mitra, S. Bhatia, Utilizing context in generative bayesian models for linked corpus, in: AAAI '10, 2010, pp. –1–1.
- [54] X. Zhang, P. Mitra, Learning topical transition probabilities in click through data with regression models, in: Proceedings of the 13th International Workshop on the Web and Databases, WebDB '10, ACM, New York, NY, USA, 2010, pp. 11:1–11:6. doi:10.1145/1859127.1859142.
- [55] R. Krestel, P. Fankhauser, W. Nejdl, Latent dirichlet allocation for tag recommendation, in: Proceedings of the third ACM conference on Recommender systems, RecSys '09, ACM, New York, NY, USA, 2009, pp. 61–68. doi:10.1145/1639714.1639726.
- [56] A. Asuncion, M. Welling, P. Smyth, Y. W. Teh, On smoothing and inference for topic models, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, AUAI Press, Arlington, Virginia, United States, 2009, pp. 27–34.
- [57] C. D. Manning, P. Raghavan, H. Schtze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.
- [58] S. Zelikovitz, H. Hirsh, Improving short text classification using unlabeled background knowledge to assess document similarity, in: Proceedings of the Seventeenth International Conference on Machine Learning, 2000, pp. 1183–1190.
- [59] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American society for information science 41 (6) (1990) 391–407.
- [60] A. Blum, T. Mitchell, Combining labeled and unlabeled data with cotraining, in: Proceedings of the eleventh annual conference on Computational learning theory, COLT '98, 1998, pp. 92–100.