Automatic Classification of Forum Posts: A Finnish Online Health Discussion Forum Case

O. Gencoglu

BioMediTech Institute and Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, Tampere, Finland

Abstract— **Online health discussion forums play a key role in accessing, distributing and exchanging health information at an individual and societal level. Due to their free nature, using and regulating these forums require substantial amount of manual effort. In this study, we propose a computational approach, i.e., a machine learning framework, in order to categorize the messages from Finland's largest online health discussion forum into 16 categories. An accuracy of 70.8% was obtained with a Naïve Bayes classifier, applied on term frequency-inverse document frequency features.**

Keywords— **machine learning, natural language processing, online discussion forum, social media, topic classification.**

I. Introduction

Social media is one of the significant aspects of the current e-health ecosystem. Online health information seekers use the Internet and social media for several reasons, e.g., researching what other consumers say about medication or treatment, researching other consumers' knowledge and experience, learning skills and gaining knowledge to manage a condition, getting emotional support, building awareness, and sharing knowledge [1]. Common platforms used by online health information consumers include blogs, wikis, social networks, live chat rooms, video-sharing websites, podcasts, online forums and message boards [2]. Social media use in healthcare is shown to have effects on patients such as enhanced psychological well-being and improved self-management and control [3]. On the other hand, addiction to social media, loss of privacy, and being targeted for promotion are also shown to be part of possible effects [3].

Online health discussion forums, while being prominent in online health communication, require governing and regulation in order to be efficient and successful due to the large amounts of unstructured information. Many online discussion forums have categorical separation of discussion topics as well as subtopics, in order to provide orderly means of communication to their users. Therefore, relevant categorization of a new message posted by a user has to either rely on user's judgment of the appropriate category or manual assignment and correction by the forum administration. In this context, employing a machine learning based topic classification system can improve the quality of the online health discussion forum by assisting both users and administrators.

In [4], posts from a smoking cessation forum are classified using a Naïve Bayes (NB) classifier. Similarly, a binary NB classifier is trained with bag-of-words features in [5] in order to classify the questions in WebMD diabetes community as important or not. In [6] a rule-based classification framework is proposed to categorize users intent of posting contents into 4 categories. In [7], handcrafted text features are extracted from online cancer survivor community posts and several machine learning algorithms are applied on the data, resulting in up to 79.2% accuracy in classifying the sentiment.

II. Methods

A. Dataset

The dataset used for this work has been extracted from Finland's largest online discussion forum with 1,400,000 weekly users, Suomi24 [8, 9]. The discussion forum consists of publicly-available, user-generated discussions that are grouped based on contents, such as entertainment, hobbies, travel, and health. Users, being anonymous, can start their own discussions or contribute to existing discussions. In this study, the forum data was retrieved from a structured database, accommodated by the service provider. The license of the database, in compliance with copyright agreements by World Intellectual Property Organization, grants the right to use and make copies of the corpus for educational, teaching and research purposes [10].

The dataset contains 352,725 posts in the *Health* category which divides into 16 sub-categories, namely ”ask your health questions”, ”birth control”, ”decease and mourning”, ”diseases”, ”drugs and addictions”, ”general health”, ”healthcare”, ”healthcare services”, ”medicines”, ”men's health”, ”mental health and wellbeing”, ”oral health”, ”plastic surgery”, ”senses (sensory organs)”, ”weight control”, and ”women's health”. The comments and discussion under the first post are not included for the analysis, i.e, only the titles and first messages (usually a question) are used for training and validating the algorithms. The distribution of number of observations among the 16 categories are not uniform, i.e, dataset holds class-inbalance. The number of messages from different categories can be examined from Table 1.

Titles and posts contain 2.6 and 75.9 words on average, respectively. The median values of word counts are 2 for titles and 49 for posts. Table 1 also shows the mean and median values of word counts for all categories.

Table 1: Number of observations, mean and median word counts for different categories.

		Word Count	
Category	Number of Posts	Mean	Median
Ask your health questions	6,012	74.5	57
Birth control	13,213	58.6	47
Decease and mourning	4,108	94.3	62
Diseases	86,035	76.4	52
Drugs and addictions	34,346	68.6	35
General health	16,415	69.2	49
Healthcare	15,461	61.0	34
Healthcare services	236	61.5	35
Medicines	9,743	52.5	36
Men’s health	4,067	55.8	36
Mental health & wellbeing	70,017	103.0	66
Oral health	10,959	59.3	43
Plastic surgery	6,123	55.6	39
Sensory organs	8,164	64.5	46
Weight control	49,257	69.1	47
Women’s health	18,569	69.5	55

B. Preprocessing

For each post in the forum dataset, the following preprocessing steps have been executed with the given order:

1. Title and message are merged with a whitespace in between.
2. All text is converted to lowercase letters/characters.
3. A whitespace is added after each . or , unless it is already there.
4. Any number of consecutive whitespace characters are transformed into a single whitespace character.
5. All urls are removed.
6. Word stemming is applied on each word with the help of Finnish language lexical database [11].

C. Feature Extraction

Term frequency - inverse document frequency (tf-idf) features are extracted from each observation in the dataset [12]. An n-gram range of [1, 2] (inclusive) are used for the feature extractor for certain runs too. With only 1-grams, the tf-idf features result in a sparse feature matrix of 352,725 rows and 1,037,221 columns, with only 0.00632% of the elements being non-zero. When 2-grams are also included, the number of features increase to 11,757,266 with non-zero elements corresponding to 0.00122% of the total.

D. Classification

Three different classifiers, namely Bernoulli NB, Multinomial NB and online passive-agressive classifier are tested in a 10-fold cross-validation (CV) manner.

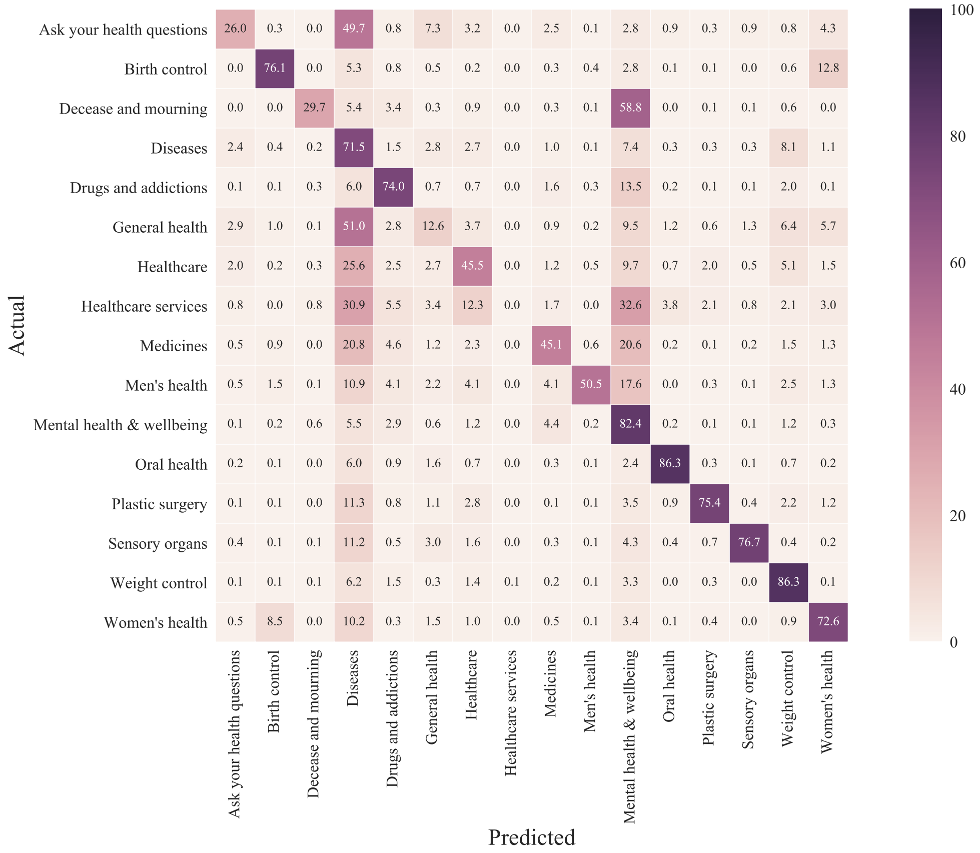
Hyper-parameters for the classifiers, e.g., loss function, regularization coefficient, decision to learn class priors or not, are selected with a grid search over a hyper-parameter space in a 10-fold cross-validation fashion as well. The final hyper-parameter values for each classifier are set to the ones that reach the highest accuracies in the CV.

Table 2: Accuracies of different algorithms and set of hyper-parameters corresponding to the best performance on 10-fold CV.

Algorithm	Best Hyper-parameters	Accuracy (%)
Passive-aggressive	C=0.03, loss = ’hinge’	68.4
Multinomial NB	alpha=0.01, prior fit=True	
without preprocessing		70.5
with preprocessing		70.8
with preprocessing + 2-grams		67.1
Bernoulli NB	alpha=0.3, prior fit=False	60.7

III. Results and Discussion

The overall classification accuracies obtained by the 10-fold CV are reported in Table 2 (with preprocessing and word tf-idf features). Highest accuracy, 70.8%, was reached by Multinomial Naive Bayes classifier which used only word tf-idf features extracted from preprocessed data. Preprocessing added 0.3% increase to the accuracy, on the other hand, extracting 2-grams in addition to 1-grams reduced the classification accuracy.



	Ask your health questions	Birth control	Decease and mourning	Diseases	Drugs and addictions	General health	Healthcare	Healthcare services	Medicines	Men's health	Mental health & wellbeing	Oral health	Plastic surgery	Sensory organs	Weight control	Women's health
Ask your health questions	26.0	0.3	0.0	49.7	0.8	7.3	3.2	0.0	2.5	0.1	2.8	0.9	0.3	0.9	0.8	4.3
Birth control	0.0	76.1	0.0	5.3	0.8	0.5	0.2	0.0	0.3	0.4	2.8	0.1	0.1	0.0	0.6	12.8
Decease and mourning	0.0	0.0	29.7	5.4	3.4	0.3	0.9	0.0	0.3	0.1	58.8	0.0	0.1	0.1	0.6	0.0
Diseases	2.4	0.4	0.2	71.5	1.5	2.8	2.7	0.0	1.0	0.1	7.4	0.3	0.3	0.3	8.1	1.1
Drugs and addictions	0.1	0.1	0.3	6.0	74.0	0.7	0.7	0.0	1.6	0.3	13.5	0.2	0.1	0.1	2.0	0.1
General health	2.9	1.0	0.1	51.0	2.8	12.6	3.7	0.0	0.9	0.2	9.5	1.2	0.6	1.3	6.4	5.7
Healthcare	2.0	0.2	0.3	25.6	2.5	2.7	45.5	0.0	1.2	0.5	9.7	0.7	2.0	0.5	5.1	1.5
Healthcare services	0.8	0.0	0.8	30.9	5.5	3.4	12.3	0.0	1.7	0.0	32.6	3.8	2.1	0.8	2.1	3.0
Medicines	0.5	0.9	0.0	20.8	4.6	1.2	2.3	0.0	45.1	0.6	20.6	0.2	0.1	0.2	1.5	1.3
Men's health	0.5	1.5	0.1	10.9	4.1	2.2	4.1	0.0	4.1	50.5	17.6	0.0	0.3	0.1	2.5	1.3
Mental health & wellbeing	0.1	0.2	0.6	5.5	2.9	0.6	1.2	0.0	4.4	0.2	82.4	0.2	0.1	0.1	1.2	0.3
Oral health	0.2	0.1	0.0	6.0	0.9	1.6	0.7	0.0	0.3	0.1	2.4	86.3	0.3	0.1	0.7	0.2
Plastic surgery	0.1	0.1	0.0	11.3	0.8	1.1	2.8	0.0	0.1	0.1	3.5	0.9	75.4	0.4	2.2	1.2
Sensory organs	0.4	0.1	0.1	11.2	0.5	3.0	1.6	0.0	0.3	0.1	4.3	0.4	0.7	76.7	0.4	0.2
Weight control	0.1	0.1	0.1	6.2	1.5	0.3	1.4	0.1	0.2	0.1	3.3	0.0	0.3	0.0	86.3	0.1
Women's health	0.5	8.5	0.0	10.2	0.3	1.5	1.0	0.0	0.5	0.1	3.4	0.1	0.4	0.0	0.9	72.6

Fig. 1: Normalized confusion matrix displaying intra-class classification accuracies.

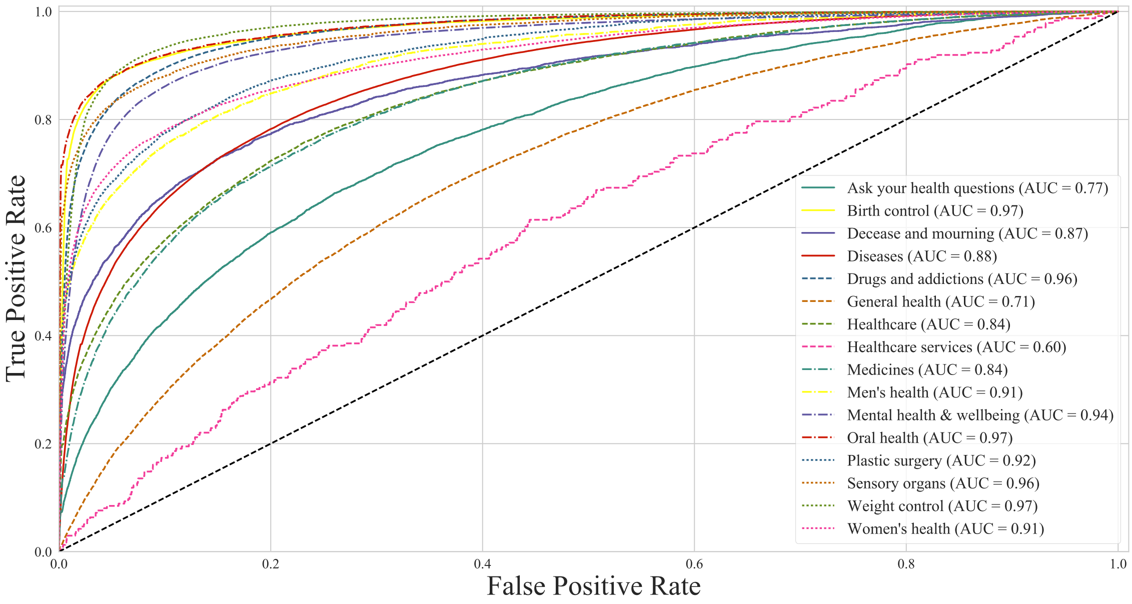


Fig. 2: ROC curves and AUC scores for each category.

Figure 1 illustrates the normalized (showing percentages) confusion matrix extracted from the run with best performance settings. The easily-classified categories were ”Oral health” and ”Weight control”, both achieving an intra-class accuracy of 86.3%. All observations of the category ”Healthcare services” were classified incorrectly. This is not an unexpected result as the number of observations in that category was very low with respect to that of other categories. A tendency from several categories to be misclassified into ”Diseases” or ”Mental health and wellbeing” classes can also be observed.

Receiver operating characteristic (ROC) curves (extracted again from the best performing setting) corresponding to each class are depicted in Figure 2. Three categories, i.e., ”Birth control”, ”Oral health”, and ”Weight control”, reached an Area Under Curve (AUC) score of 0.97.

The t-SNE [13] mapping of 2000 observations from ”Oral health” and ”Weight control” classes (randomly selected 1000 for each) into first 2 dimensions can be seen in Figure 3. Even with only 2 dimensions, certain level of separability can be observed.

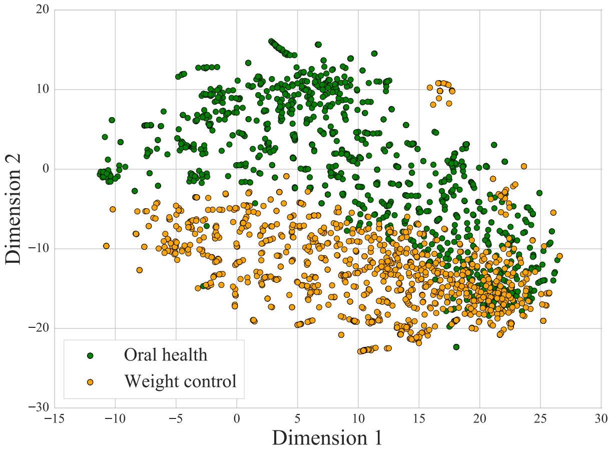


Fig. 3: t-SNE mapping of 2000 observations from ’Oral health’ and ’Weight control’ topics into 2 dimensions.

IV. Conclusion

Automatic classification of messages in online health discussion forums is valuable for ease of seeking, providing, retrieving and regulating health information. With the help of machine learning, both users and administrators of the forums can be steered towards better categorization of forum content, resulting in an enhanced experience of health information exchange. For future studies, recent topic classification approaches such as deep learning will be studied.

Conflict of Interest

Author declares no conflict of interest.

Acknowledgements

This research has been supported by a grant from Tekes - the Finnish Funding Agency for Innovation as part of Digital Health Revolution programme.

References

1. Levy M. Online health: assessing the risk and opportunity of social and one-to-one media *Jupiter Research.* 2007;2.
2. Sarasohn-Kahn Jane. The wisdom of patients: Health care meets online social media 2008.
3. Smailhodzic Edin, Hooijsma Wyanda, Boonstra Albert, Langley David J. Social media use in healthcare: A systematic review of effects on patients and on their relationship with healthcare professionals *BMC Health Services Research.* 2016;16:442.
4. Zhang Mi, Yang Christopher C. Classification of Online Health Discussions with Text and Health Features Sets in *Proceedings of AAAI International Workshop on the World Wide Web and Public Health Intelligence 2014 (W3PHI 2014)* 2014.
5. Huh Jina, Yetisgen-Yildiz Meliha, Pratt Wanda. Text classification for assisting moderators in online health communities *Journal of Biomedical Informatics.* 2013;46:998–1005.
6. Liu Jun, Shang Yanyan. Users' Continuance Participation in the Online Peer-to-peer Healthcare Community: A Text Mining Approach 2015.
7. Qiu Baojun, Zhao Kang, Mitra Prasenjit, et al. Get online support, feel better–sentiment analysis and dynamics in an online cancer survivor community in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*:274–281IEEE 2011.
8. Suomi24 http://www.suomi24.fi/, Accessed 3 March 2017, Archived by WebCite at http://www.webcitation.org/6oeDNxg8M.
9. Suomi24 profile card http://www.aller.fi/wp-content/uploads/2014/12/Suomi_24_profiilikortti.pdf, Accessed 3 March 2017, Archived by WebCite at http://www.webcitation.org/6oeE3FLCT.
10. Suomi24 corpus http://metashare.csc.fi/repository/download/b4db73da85ce11e4912c005056be118ea699d93902fa49d69b0f4d1e692dd5f1/, Accessed 3 March 2017, Archived by WebCite at http://www.webcitation.org/6oeEAXcAo.
11. Bird Steven. NLTK: the natural language toolkit in *Proceedings of the COLING/ACL on Interactive Presentation Sessions*:69–72Association for Computational Linguistics 2006.
12. Ramos Juan, others . Using tf-idf to determine word relevance in document queries in *Proceedings of the First Instructional Conference on Machine Learning* 2003.
13. Maaten Laurens van der, Hinton Geoffrey. Visualizing data using t-SNE *Journal of Machine Learning Research.* 2008;9:2579–2605.

Author: Oguzhan Gencoglu
Institute: Tampere University of Technology
Street: Korkeakoulunkatu 10
City: Tampere
Country: Finland
Email: oguzhan.gencoglu@tut.fi