Automatic extraction of personal experiences from patients' blogs: A case study in chronic obstructive pulmonary disease

Mark Greenwood[1,2], Glyn Elwyn[2], Nick Francis[2], Alun Preece[1], Irena Spasić[1]

[1]School of Computer Science & Informatics
Cardiff University
Cardiff, UK

[2]Cochrane Institute of Primary Care & Public Health
Cardiff University
Cardiff, UK

***Abstract*— People with long-term illness such as chronic obstructive pulmonary disease (COPD) often use social media to document and share information, opinions and their experiences with others. Analysing the self-reported experiences of patients shared online has the potential to help medical researchers gain insight into some of the key issues affecting patients. However, the scale of health conversation taking place online poses considerable challenges to traditional content analysis. In this paper, we present a system which automates extraction of patient statements which refer to a personal experience. We applied a crowdsourcing methodology to create a set of 1770 annotated sentences from blog posts written by COPD patients. Our machine learning approach trained on lexical features successfully extracted sentences about patient experience with 93% precision and 80% recall (F-measure: 86%). Automatic annotation of sentences about patient experience can facilitate subsequent content analysis by highlighting the most relevant sentences to this particular problem.**

Keywords—blogs; blog mining; machine learning; text processing; natural language processing; health informatics; social media

I. INTRODUCTION

User-generated content and the popularity of social media have enabled people to communicate online with potentially large audiences all over the world. This has had an impact on healthcare and how people find and share health information. NM Incite (a market researchers focusing on social media) have described an online community of hundreds of thousands of patients and carers taking part in an online discussion about a wide range of health conditions [1]. Table 1 shows statistics from their infographic *Healthcare Social Media by the Numbers* highlighting an online discussion about health where the majority of participants are living with or caring for people with these long term conditions.

TABLE I. ONLINE AUTHORS FOR THE 5 MOST PREVALENT CONDITIONS – ADAPTED FROM [1]

Condition	Condition Prevalence	Authors	Participants		
			Patient	*Carer*	*Healthcare Professionals*
Cardiovascular Disease	81m	65k	59%	39%	2%
Arthritis	52m	38k	92%	7%	1%
Asthma	24m	45k	72%	28%	0%
COPD	24m	12k	64%	36%	0%
Type 2 Diabetes	16m	24k	87%	13%	0%

There are a wide range of social media platforms which enable users to create and share content online, including social networking sites (e.g. Facebook, Google), content communities (e.g. YouTube, Instagram), collaborative projects (e.g. Wikipedia) and blogging platforms (such as Blogger, Wordpress, etc.) [2]. Walker described blogs in terms of their appearance or blogging platforms' functionality as "frequently updated websites consisting of dated entries in reverse chronological order so the most recent post appears first" [3]. Along with technical features of a blog, Winer characterises blogs in terms of their content – the website should contain the "unedited voice of a person" to qualify as a blog [4]. In other words it should be about the personal experiences or opinions of the author (or *blogger*). Blogging platforms are used to discuss health issues online. Technorati[1] hosts a manually curated blog directory which, at the time of writing, was tracking over 23,000 health related blogs[2]. These blogs are used to share information, personal opinions and experiences around health and medicine with a general audience online.

Experiences shared by patients through online media are an important resource for other people faced with similar issues. Recent surveys carried out by Pew Internet in the USA showed that 34% of internet users (25% of adults) had consulted other peoples' commentary or experience of healthcare shared online when faced with an information need [5] and that 26% had used this source of information in the past 12 months [6].

Healthcare providers also rely on patient experience to improve services. In the UK, the NHS constitution [6] states

[1] http://www.technorati.com

[2] http://technorati.com/blogs/directory/living/health/

978-0-7695-5114-2/13 $26.00 © 2013 IEEE
DOI 10.1109/CGC.2013.66

that one of the organisation's key principles is that patients, their families and carers should be involved in and consulted on decisions about their treatment and health care. The NHS has also set aims to involve service users in organisational level decisions through public consultations [7]. Patient reported outcome measures (PROM), where service users are surveyed about their treatments, are currently used to evaluate treatments and services [8].

Patient experience is also important in setting research priorities. For instance the Lind Alliance[3] is an organisation which supports bringing patients, carers and clinicians together to involve all stakeholders in setting research priorities according to patient needs.

Online patient discussions have been used to help explore patient experiences. Sillence and Mo carried out a qualitative analysis of forum posts to understand how patients make decisions [10]. Similarly, Hewitt-Taylor and Bond analysed online discussion boards to unravel patient expectations of their relationship with their physicians [11]. Blogs [12] and video blogs [13] of cancer patients have been analysed manually to better understand patients' experience of care. Understanding these information sources is essential in allowing them to be used more effectively by healthcare professionals and researchers as well as other patients.

Our work aims to enable easier, faster access to patient experiences shared through blog posts by facilitating content analysis using large-scale text mining. In this paper, we discuss our approach to automatic binary classification of sentences from patient blog posts with respect to their reference to a subjective patient experience (e.g. an expression of a personal narrative or opinion). We demonstrate our approach in a specific health domain, namely chronic obstructive pulmonary disease (COPD). COPD is described by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) as a disease which causes a progressive limitation of airflow within the lungs, which is not fully reversible [14]. It has been predicted that by 2020 COPD will be the fifth leading cause of disability and the third leading cause of death worldwide [15]. Much of COPD patients' day-to-day care is self-managed [16] and therefore takes place outside of the direct care setting. This means that patients need to become more informed about their condition and online health communities are particularly important to this patient population. COPD is therefore a useful case study for our approach.

II. METHOD

Blogs are used by authors to share their personal experiences and opinions as well as information and advice. For example,

Personal experience:

"After my discharge from hospital a couple of weeks ago I continue to monitor my condition and so far, so good with no sign of another infection."

Information:

"Some COPD patients live alone, and are in many aspects isolated."

Advice:

"A better plan might be to let your doctor know what is going on so that he or she can find a way to relieve the problem."

This paper sets out our approach to filtering sentences related to the author's personal experience from the other kinds of content.

Given the subjective and highly variable nature of the problem, we have opted for a supervised machine learning approach as opposed to a rule-based approach. Such an approach relies on a training set of annotated examples. In this case that implies the collection of relevant documents (i.e. blog posts) and manual annotation of individual sentences that refer to a subjective patient experience. Finally, it remains to select an appropriate set of features that adequately characterise subjective sentences. This section details our approach.

A. Data Collection

1) Corpus

In order to compile a text corpus, we first needed to identify relevant blog posts written by COPD patients or their carers. We used blog search engines (Google Blog Search[4], Technorati) to retrieve relevant blogs using a set of search terms related to COPD, namely *COPD*, *chronic obstructive {pulmonary|lung|airways|respiratory} disease*, *bronchitis* and *emphysema*. A total of 50 active blogs specifically related to COPD were selected initially. Given the relatively small size, these blogs were then reviewed manually. Blogs authored by patients and carers were included (17 blogs), whereas blogs authored by physicians, companies or others were excluded (14 blogs) as were marketing blogs (19 blogs). To support future applications on a larger scale, these annotations can be used to train text classification algorithms. . As we intended to automatically process the content of the blog posts and quote them elsewhere, we also discarded blogs with restrictions on content use.

In order to collect the content of the selected blogs, we used RSS feeds supplied through blogging platforms. RSS feeds provide a stream of recent posts in an XML format. Blogs without RSS support were not collected. Table 2 summarises the properties of the collected blog posts.

TABLE II. CORPUS PROPERTIES

Blogs	12
Authors	44
Blog posts collected	368 (819KB)
Average length (tokens)	461 (std dev: 402)
Post dates	2006-2012
Sentences	7955
Tokens	165042
Distinct tokens	13861
Mean sentence length (tokens)	20.7 (std dev: 14.5)

[3] http://www.lindalliance.org/

[4] http://www.google.com/blogsearch

The collected documents were linguistically pre-processed using the Stanford Part-Of-Speech Tagger [17], which, along with part-of-speech tagging, provides sentence splitting and tokenization of the text data.

2) Annotation

In order to create a training dataset of manually annotated sentences referring to a subjective patient experience, we used crowdsourcing to collect multiple opinions from a wide range of people over the Internet. Crowdsourcing is a method of problem solving or content creation through distributed participation of many individuals following an open call [18]. The use of the Internet as a crowdsourcing medium bridges the physical gap and allows a large number of participants to be reached. An open call for participation enabled us to collect opinions from various stakeholder groups. Interpreting sentences as pertaining to a subjective experience does not require specialist knowledge and using annotations from a spread of stakeholder groups reduces interpretation bias. However, anonymous data collection online is prone to gaming and malicious behaviour, which increases the probability of poor quality annotations [19, 20, 21]. In order to address this risk each sentence was annotated by more than one user. Firstly, this allowed us to assess the quality of the annotations using inter-annotator agreement analysis. Secondly, applying a majority vote method, we could select the annotations with the highest agreement for training and testing our classifiers.

As multiple annotations were required for each sentence we reduce the set of sentences, selecting a subset of 100 blog posts (from the set of 368) and including the 1770 sentences in them containing more than 10 tokens. Using a smaller set would allow this overlap in users' annotations with the response expected to the open call.

We implemented a web-based annotation tool[5] to display 20 random sentences (from the set of 1770) to each participant at any one time and allow them to annotate each sentence as relating to a subjective experience or not. Each sentence was displayed within the context it appeared in originally (including the preceding and proceeding sentence) in order to provide context for the purpose of annotation (see Figure 1).

Fig. 1. Annotation site screenshot

Manual annotations, summarised in Table 3, were collected in three phases:

1. Small scale pilot (12 participants) validating online exercise.
2. Call to participate through mailing lists, department newsletters and social networking sites (Facebook, Twitter).
3. Repeat invitations, plus recruitment through Amazon Mechanical Turk crowdsourcing platform[6].

Each participant was asked to describe their stake relating to this task (e.g. patient, carer, student, medical professional, etc.). Information on participants' self-identified stake is described in Figure 2.

TABLE III. ANNOTATIONS PROPERTIES

Blog posts	100 (233KB)
Sentences with >10 tokens	1770
Annotators	286
Completers (annotated all 20 sentences)	226
Total annotations	4745
Average annotations per sentence	2.68 (std dev: 0.5)

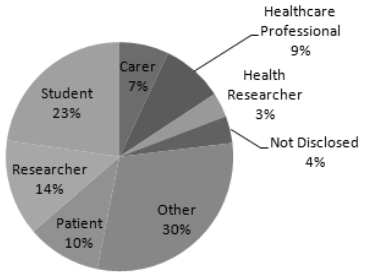

Fig. 2. Annotators' stake

Figure 2 shows that our call for participants reached a range of stakeholders including inside academia (researchers, students), healthcare (healthcare professionals and researchers) as well as patients and carers. This spread allows our dataset to represent a number of interpretations of the blog post sentences.

We used Krippendorff's alpha coefficient [22] to rate the inter-annotator agreement. The coefficient α=0.55 implied less than perfect agreement between annotators, but greater than that expected by chance (α=0).

Table 4 describes the annotations in terms of inter-annotator agreement. The majority (1247 out of 1770) of sentence annotations were agreed on unanimously between all annotators to whom they were shown. All but 135 sentences had a majority vote in favour of a single annotation.

[5] http://users.cs.cf.ac.uk/M.A.Greenwood/annotation

[6] http://www.mturk.com

TABLE IV. SENTENCES BY ANNOTATION AGREEMENT

Agreement	Sentences
No agreement	135
Majority agreement (2/3, 3/4)	388
Total agreement	1247

The annotated data was used to create two datasets. The first contains only sentences where the annotation was agreed on unanimously (i.e. agreement = 100%, *n*=1247) and the second, where there was a majority in favour of one annotation (i.e. agreement >50%, *n*= 1635). Using agreement as a measure of annotation quality, we will train and test classifiers on each dataset independently in order to assess its impact on classification performance.

B. Feature space

Our approach focused on a token-level representation of each sentence within the dataset. nouns, verbs, adjectives and adverbs, were found in each sentence and generalised according to their meaning using WordNet [23] synsets. WordNet is a lexical database providing information about general English words. Synsets allow grouping of words with the same meaning (see Figure 3) and were used to group similar tokens (i.e. 'disbelieving' and 'sceptical') into more general features thereby creating a more useful model for generalisation.

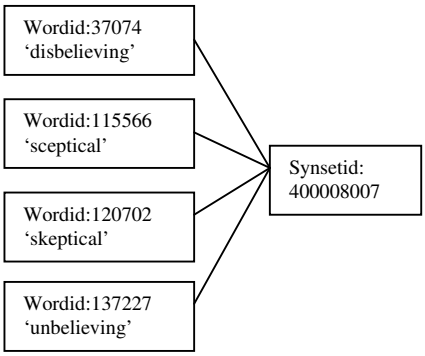

Fig. 3. WordNet synset mapping

Personal pronouns used in sentences were also included in the feature set. Pronouns were grouped in to three classes – first person ('me', 'I', etc.), possessive ('my', 'mine', etc.) and third person ('you', 'them', 'they', etc.). When discussing a personal experience, statements in the first person would be expected intuitively. Pronouns would therefore be a potentially valuable feature for classifying experiential sentences. A summary of the feature groups and their 5371 features can be found in Table 5.

TABLE V. FEATURE SPACE

Feature	Description
Length	Length of the sentence in tokens
Pronouns	Relative frequency of personal pronouns present separated into three classes 1. First person ('I', 'me', etc.) 2. Possessive ('my', 'mine', etc.) 3. Third person ('them', 'they', etc.)
Nouns	Relative frequency of noun tokens, grouped by synset (1998)
Verbs	Relative frequency of verb tokens, grouped by synset (713)
Adjectives	Relative frequency of adjectives, grouped by synset (644)
Adverbs	Relative frequency of noun tokens, grouped by synset (212)

C. Feature selection

The discriminative power of each feature was assessed using information gain analysis [24]. Information gain values represent the weight of information held by a feature regarding a class. It is one indication of how useful features are for discriminating between classes. The general form of Information gain for nominal classes is computed as:

$$\text{Information Gain}(\text{Class}, \text{Attribute}) = H(\text{Class}) - H(\text{Class}|\text{Attribute})$$

where H represents the information entropy – a measure of uncertainty about a random variable.

Features were then ranked in order of the information gain values associated with the annotations (i.e. experiential sentence or not). The top 10 features according to this analysis are shown in Table 6 and a summary of the results of the analysis in Table 7.

TABLE VI. TOP 10 FEATURES FROM INFORMATION GAIN ANALYSIS

Rank	Feature (Info. Gain value)	
	Agreement 100%	*Agreement >50%*
1	First person pronoun (0.48)	First person pronoun (0.358)
2	'have' (0.047)	'have' (0.033)
3	'disease' (0.039)	'disease' (0.031)
4	'lung' (0.032)	'patient' (0.023)
5	'patient' (0.031)	'chronic' (0.02)
6	'get' (0.027)	'lung' (0.019)
7	'chronic'(0.027)	'last' (0.018)
8	'last' (0.021)	'so' (0.018)
9	'symptom' (0.18)	'get' (0.015)
10	'good' (0.18)	'move' (0.015)

TABLE VII. NUMBER OF POTENTIALLY USEFUL FEATURES BY FEATURE GROUP

	Number of features where Information Gain>0	
	Agreement 100%	*Agreement >50%*
Nouns	76	80
Verbs	30	33
Adjectives	19	19
Adverbs	14	12
Pronouns	2	2
Total	141	146

Table 6 shows that the relative frequency of first person pronouns in a sentence exhibits the most discriminative power as intuitively expected. Tokens from other feature groups (nouns, verbs, adjectives and adverbs) are also highly ranked, but with much lower values, showing that they are not expected to be useful for this classification task.

D. Model Building

To evaluate the features chosen to represent sentences in our dataset, models were trained on various combinations of features from the set described in section 2.3. Subsets of our overall feature set (i.e. nouns, verbs, adjectives, adverbs and pronouns) were evaluated separately as well as using them all together. For each subset the ranked set of token-features with non-zero information gain values were used (see Table 7) to train a Naive Bayes classifier. Naive Bayes classifiers [25] estimate the probability of a hypothesis based on previous evidence and Bayes' theorem:

$$P(h|e) = \frac{P(h)P(e|h)}{P(e)}$$

During the training phase, the training data is used to estimate parameters for the probability distributions. This can then be used to estimate the probability of a hypothesis on previously unseen data. A Naive Bayes classifier was chosen as it performs well with relatively small training sets [25] as shown in a previous text classification study [26]. In our experiments, it also performed better than alternative approaches such as Support Vector Machines and J48 decision trees. Classification models were built and evaluated using 10-fold cross-validation in the Weka [27] machine learning package, version 3.6.

III. RESULTS

The classifiers were trained and tested on the two overlapping sets of sentences using the selected features in combination and isolation in order to evaluate their appropriateness for this classification task. As the results of the information gain analysis (Table 6) show, the pronouns used in the sentence expected to be most informative. The results of our experiments are shown in Table 8 and Table 9.

TABLE VIII. NAIVE BAYES CLASSIFIER RESULTS (AGREEMENT > 50%)

	No. of Features	Precision	Recall	F-Score	Kappa
Nouns	80	0.643	0.942	0.764	0.3253
Verbs	33	0.64	0.818	0.718	0.2703
Adjectives	19	0.579	**0.991**	0.731	0.1291
Adverbs	13	0.784	0.288	0.422	0.1804
Pronouns	2	**0.902**	0.728	**0.806**	**0.6201**
All	147	0.672	0.94	0.784	0.4024

TABLE IX. NAIVE BAYES CLASSIFIER RESULTS (AGREEMENT =100%)

	No. of Features	Precision	Recall	F-Score	Kappa
Nouns	76	0.449	0.6	0.767	0.4333
Verbs	30	0.666	0.656	0.78	0.3583
Adjectives	19	0.6	**0.993**	0.748	0.1695
Adverbs	19	0.825	0.311	0.452	0.2102
Pronouns	2	**0.933**	0.803	**0.864**	**0.7174**
All	141	0.719	0.958	0.822	0.5064

Tables 8 and 9 show the results achieved by the different feature sets on the two datasets described. With both unanimous and majority agreement datasets, the pronoun features performed the best when classifying experiential sentences. First person and possessive pronouns were used to achieve over 90% precision in both datasets, and 72-80% recall. The chance-corrected kappa scores for both the majority and unanimous datasets show that the pronoun features performed much better than the random classifier baseline (0.62 and 0.72 respectively). The Kappa scores attached to each outcome compare the results achieved to picking classes at random. A Kappa score of 0 signifies a result no better than random while a score of 1 signifies perfect agreement.

The combined features achieved greater recall at the expense of precision. Kappa agreement values showed results closer to that expected by chance (0.40 for the majority dataset, 0.51 for the unanimous dataset).

In summary our results show that sentences relating to patient experience can be automatically extracted from patient accounts shared online with good confidence. They indicate that the frequency of first-person or possessive pronouns is the most appropriate features of the sets tested. The high precision of pronouns (90-93%), but relatively low recall (72-80%) indicates this feature set should be expanded in order to increase coverage, but the general classes of token used here were largely unsuccessful. Adding more information to the textual data may add value to the features selected. Utilising emotional lexicons, such as WordNetAffect [28] and SentiWordNet [29] will allow us to generalise tokens in terms of the emotion they express, which could be a useful feature here. Phrase-level features, rather than token-level may also increase the information load.

IV. CONCLUSION

Understanding how patients use social media to communicate information and experiences of illness and how this information may be better used to improve care is an important health informatics challenge. Our work aims to make accessing patient experiences shared online faster and easier through enabling automatic interpretation.

The agreement achieved during the crowdsourcing task show that patient experience is something a wide range of people, in and outside of healthcare, can relate to and interpret. The results of our automatic classification experiments show that these sentences can be automatically identified, providing easier access to these accounts to healthcare researchers.

In our experiments first-person and possessive pronouns were found to be the most effective predictor when classifying

sentences as pertaining to a patient's experience. This is an intuitive result as personal accounts would be expected to be self-referential.

The next stage in this work is to evaluate to what degree automated text classification can facilitate traditional qualitative research about patient experience. Presenting this information to healthcare researchers for further interpretation and utilisation can be approached using open standards for qualitative data exchange, such as QuDex, created by the Data Exchange Tools project (DExT) [33]. Future developments will include information extraction to further support automatic interpretation of data in this domain. Relevant domain-specific concepts (e.g. medications, treatments or healthcare professionals) will be identified in text by using large dictionaries such as the Unified Medical Language System [30] or automatic term recognition approaches such as FlexiTerm [31]. Emotions expressed by the author could also be extracted using sentiment analysis [32]. Supporting the automatic analysis of patient experiences shared online is an important step to helping healthcare researchers make use of this important information source. By enabling faster access to the most relevant information, patient opinions on healthcare can directly influence its future.

REFERENCES

[1] NM Incite. (2011). Healthcare Social Media by the Numbers. Retrieved July 16, 2012, from http://nmincite.com/healthcare-social-media-by-the-numbers/

[2] Kaplan, AM., Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53, pp. 59-68.

[3] Walker, J. D. (2003). jill/txt: final version of weblog definition. Retrieved April 24, 2013, from http://jilltxt.net/archives/blog_theorising/final_version_of_weblog_definition.html

[4] Winer, D. (2003). Harvard Weblogs: What makes a weblog a weblog? Retrieved April 3, 2013, from http://blogs.law.harvard.edu/whatmakesaweblogaweblog.

[5] Fox, S. (2011). The Social Life of Health Information, 2011. Retrieved June 14, 2011, from http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx

[6] Fox, S., & Duggan, M. (2013). *Health Online 2013*. Retrieved from http://www.pewinternet.org/Reports/2013/Health-online.aspx

[7] Department of Health. (2013). The NHS Constitution for England. Department of Health, London, UK

[8] Department of Health. (2006). Our health, our care, our say: a new direction for community services (White Paper). Department of Health, London, UK

[9] Marshall, S., Haywood, K., & Fitzpatrick, R. (2006). Impact of patient-reported outcome measures on routine practice: a structured review. *Journal of evaluation in clinical practice*, *12*(5), 559–68. doi:10.1111/j.1365-2753.2006.00650.x

[10] Sillence, E., & Mo, P. K. H. (2012). Communicating health decisions: an analysis of messages posted to online prostate cancer forums. *Health expectations : an international journal of public participation in health care and health policy*.

[11] Hewitt-Taylor, J., & Bond, C. S. (2012). What E-patients Want From the Doctor-Patient Relationship: Content Analysis of Posts on Discussion Boards. *Journal of medical Internet research*, *14*(6), e155. doi:10.2196/jmir.2068

[12] Keim-Malpass, J., & Steeves, R. H. (2012). Talking with death at a diner: young women's online narratives of cancer. *Oncology nursing forum*, *39*(4), 373–8, 406. doi:10.1188/12.ONF.373-378l

[13] Chou, W.-Y. S., Hunt, Y., Folkers, A., & Augustson, E. (2011). Cancer survivorship in the age of YouTube and social media: a narrative analysis. *Journal of medical Internet research*, *13*(1), e7. doi:10.2196/jmir.1569

[14] GOLD. (2010). *Pocket Guide to COPD Diagnosis, Management and Prevention: A Guide for Health Care Professionals*. Retrieved from http://www.goldcopd.org/uploads/users/files/GOLD_Pocket_2010Mar31.pdf

[15] MacNee, W., & Rennard, S. (2004). *Fast Facts: Chronic Obstructive Pulmonary Disease*. Health Press Limited.

[16] NICE. (2004). CHRONIC OBSTRUCTIVE PULMONARY DISEASE - National clinical guideline on management of chronic obstructive pulmonary disease in adults in primary and secondary care. *Thorax*, *59*((Suppl I)), 1–232.

[17] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03* (Vol. 1, pp. 173–180). Morristown, NJ, USA. doi:10.3115/1073445.1073478

[18] Brabham, D. C. (2008). Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies*, *14*(1), 75–90. doi:10.1177/1354856507084420

[19] Vuurens, J., de Vries, A. P., & Eickhoff, C. (2011). How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11) (pp. 21-26).

[20] Zhu, D., & Carterette, B. (2010). An analysis of assessor behavior in crowdsourced preference judgments. In SIGIR 2010 workshop on crowdsourcing for search evaluation (pp. 17-20).

[21] Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning From Crowds. *The Journal of Machine Learning Research*, *11*, 1297–1297–1322–1322. Retrieved from http://dl.acm.org/citation.cfm?id=1756006.1859894

[22] Krippendorff, K. (1970). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, *30*(1), 61–70. doi:10.1177/001316447003000105

[23] Princeton University "About WordNet". WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>

[24] Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE- (pp. 412-420). MORGAN KAUFMANN PUBLISHERS, INC..

[25] Domingos P, Pazzani M. (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning. 29:103–37.

[26] Spasić, I., Burnap, P., Greenwood, M., & Arribas-Ayllon, M. (2012). A naïve bayes approach to classifying topics in suicide notes. *Biomedical informatics insights*, *5*(Suppl. 1), 87–97. doi:10.4137/BII.S8945

[27] Hall, M., Eibe, F., Holmes, G., Pfahringer,B.,Reutemann, P., Witten, I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, Vol. 11, Issue 1.*

[28] Valitutti, R. & Stock, O., 2004. Developing Affective Lexical Resources. *Psychology*, 2, pp.61-83. Available at: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.2710.

[29] Esuli, A. & Sebastiani, F., 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of LREC*.

[30] NLM, 2012. Unified Medical Language System. Available at: http://www.nlm.nih.gov/research/umls/ [Accessed July 17, 2012].

[31] Spasić, I., Greenwood, M., Preece, M., Francis, N. & Elwyn G. (2013). FlexiTerm: A flexible term recognition method. *Journal of biomedical semantics*, to appear

[32] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. doi:10.1561/1500000011

[33] DExT. (2013). QuDex schema. Avaiable at: http://dext.data-archive.ac.uk/schema/ [Accessed: 09/06/2013]