A Supervised Machine Learning Study of Online Discussion Forums about Type-2 Diabetes

Jonathan-Raphael Reichert[1], Klaus Langholz Kristensen[1], Raghava Rao Mukkamala[1,2], Ravi Vatrapu[1,2]
{rrm.digi,vatrapu}@cbs.dk,
[1]Centre for Business Data Analytics, Dept. of Digitalization, Copenhagen Business School, Denmark
[2] Mobile Technology Laboratory, Westerdals Oslo School of Arts, Communication and Technology, Norway

***Abstract*—As an instance of online communities, online diabetes discussion forums mirror these characteristics and seem to track the growing impact of diabetes on individuals around the world. In this paper, we first systematically collected texts from online discussion forums about diabetes and then applied supervised machine learning techniques to analyze the online conversations. In order to analyse these online textual conversations, we have chosen four domain specific models (*Emotions*, *Sentiment*, *Personality Traits* and *Patient Journey*). As part of text classification, we employed the ensemble learning method by using 5 different supervised machine learning algorithms to build a set of text classifiers by using the voting method to predict most probable label for a given textual conversation from the online discussion forums. Our findings show that there is a high amount of trust expressed by a subset of users and these users play a vital role in supporting other users of the online discussion forums about diabetes.**

I. Introduction

It has become common for people in western societies to adopt online behaviour in their everyday life and people are increasingly interested in conversing topics ranging from politics to eating habits and health status. The latter has traditionally been something very private and something that we would only talk with our family and friends or with a doctor. A study from 2013 shows that in USA, one of the world's most technologically advanced countries, 72% of internet users have looked for health information online within the past year [1].

Nowadays, many online communities are focussed on health-related issues. One such health-related issue is diabetes and diabetic patients are joining online communities to discuss everything related to their disease, from exercise and nutrition, to treatment and diagnosis. In a world where such communities clearly hold an abundance of information, organisations are becoming increasingly aware of the value hidden within these communities. However, many organisations in general, public helath organizations in particular, are struggling with challenges of extracting valuable insights from the data, so that they can turn it into actionable insights. Many organisations still only use digital and social media channels to communicate information about themselves. Research [2] shows that organisations are not utilising social platforms sufficiently and moreover they do not encourage patients to share personal experiences that might benefit other users. The main challenge for organisations is about analysing the data shared by users of various online communities, as it tends to be unstructured and messy. Recent literature has shown that researchers are trying to develop ways to apply machine learning for handling social media data [3]–[6]. Machine learning is amongst the newer technologies used to find patterns and insights in large and messy datasets.

In this paper, we argue that machine learning is an effective method to process data from online diabetes conversations, and is relevant in the application of a larger dataset to potentially extract new insights that can benefit patients and healthcare professionals. Furthermore, these insights, and the models we built, can be applicable for large pharmaceutical companies to incorporate in their holistic view of strategizing and communicating. We believe that a better understanding of online communities will eventually benefit organisations with the possibility of a more relevant and better targeted communication than before.

Type 2 diabetes was once known as adult-onset diabetes, but those days are long gone and nowadays even young children are diagnosed with type 2 [7]. Diabetes has become an issue of almost pandemic proportions, and there is a need to challenge this disease in every possible way. For patients, healthcare systems, and pharmaceutical companies, it is of great importance that researchers around the world develop new ways to either prevent the disease, or at least enlighten people about the ramifications that come with it. In this way, people diagnosed with diabetes might have a chance for an improved life with diabetes. Observing that the majority of research about diabetes is either clinical or in an offline setting, we believe there is much to learn from investigating these online conversations.

The main focus of this paper is to explore how supervised machine learning techniques can be applied to diabetes conversations and possibly to extract valuable insights from the user-generated content. Furthermore, these insights might be of helpful to those who are affected by diabetes, from patients and doctors, to pharmaceutical companies and public health organisations. Consequently, these valuable insights could be utilised to improve the various communicational aspects that revolve around online behaviour of patients, and therefore relevant to the corporate communication context as well.

We are interested in investigating how healthcare sector can leverage online forum conversations to get a better understanding of diabetes-related conversations. We believe that conversations data from online communities can be leveraged

978-1-5090-6704-6/17/$31.00 ©2017 IEEE

strategically and become an important part of an outside-in approach to corporate communication. This could ultimately optimise various strategic aspects of healthcare organisations or pharmaceutical companies, and thus be relevant for patients in their presumed pursuit of better healthcare solutions. Thus, we believe that investigating online communities and topic-oriented means of extracting online data can become an essential part of learning how to improve the healthcare sector. Our primary research question is

> *Which insights can be derived from online forum conversations about type 2 diabetes, and how can such insights be used to optimise healthcare communication and services to benefit diabetes patients, pharmaceutical companies, and healthcare organisations?*

To answer the research question, we investigated the following sub-questions:

1) Which main interests and challenges may be found from interpreting online patient conversations about diabetes?
2) To what extent can machine learning, through text classification, support or clarify qualitative insights?
3) In what way can insights from online conversations about type 2 diabetes lead to strategic recommendations?

The rest of the paper is organised as follows. In Sec. II, we will discuss about main theoretical concepts behind our research work and also mention the related studies briefly. Research methodology for this work is explained in Sec. III and results of our data analysis will be discussed in Sec. IV. Finally we will discuss and interpret the results and conclude the paper in Sec. V.

II. Conceptual Framework and Related Work

A. Corporate Communcation

Organisations have started to explore market insights and customer relations as part of their external communications and strategies, rather than simply pushing products onto the market [8]. This new outside-in approach [8] focuses on profiting from customers, by considering them as valuable assets to reach new opportunities. Market insights are thoroughly analysed data, and accurate reflections from the market are extracted through market intelligence and analytic tools built for the purpose [8]. With technological development, more and more tools are made available for organisations, in areas such as data mining, cloud services, in-depth surveys, dashboards, statistical software, etc. These tools may be leveraged in an outside-in approach to eliminate communication asymmetries and employ a user-centred approach [8]. However, it is important to emphasise that not all communication is or needs to be performed as two-way communication. Most digital assets that we see today do not talk with, but to customers, patients, or users. Many of these assets, be it corporate websites, branded content, or unbranded domains, are created to communicate to one or more stakeholders. This one-way communication may or may not be created based on user insights. If this is not the case, the user might lose interest, leave, and may never come back. Good, interesting, and strategically well-thought content, which is built on user insights and hence users' interest and needs, can still function as impactful one-way communication. One-way communication may still be relevant and can potentially lead to behavioural change. However, the one-way communication will have to fight with other sources of information and be interpreted in accordance with the users' existing knowledge, interest, and contextual setting. Hence, persuasion is still a factor to be taken seriously.

Organisations could benefit from adopting a pull-strategy in combination with the traditional push-strategy. Merely pushing products towards users with the expectation that the product offered will be valuable and attractive enough for the target audience is a challenging act in a competitive market. Moreover, when consumers become more knowledgeable and learn to understand the market and its offerings, organisations may want to distinguish themselves by having a relational approach [8]. With this knowledge, the organisation can try to adapt their offerings and communications towards audience needs and wishes [8]. Especially in a time when social media has become popular, customer insights leveraged from social media sites give organisations the opportunity to adapt their communication to fit the needs and expectations of consumers. This outside-in approach is what drives our methodological choice and analytical design, as we aim to assist the healthcare sector in improving their existing communications with diabetes patients.

B. Collective Value Creation in Online Communities

Computer-mediated communication affords new ways to interact with technology and with other individuals. One of the interesting aspects of this is the vast amount of digital traces these interactions leave behind. These latent sources of information are available via online communities and have sparked an interest in modern researchers and marketers. In [9] Johnson and Ambrose questioned the postmodern thesis that individuals are individualists. Instead, they argued that individuals are social entities that gather in fluid networks for social interaction [9]. These types of networks have existed throughout history, but have come to manifest themselves as virtual communities in the postmodern society. The popularity of online communities has emerged due to the socio-technical fit between participants' needs and technology's ability to meet these needs [9]. In the healthcare system, these communities form to fill voids that exist due to the professional healthcare systems' inability to meet up the needs and necessities of patients [9]. If patients do not get sufficient information or support from their healthcare professionals, social platforms can act as a substitute to fulfil this need. In recent studies [10], Wenger et. al. investigated online communities as a common place for learning and knowledge sharing. Wenger et al. point out that what radically differs online communities from traditional communities is the fact that communication is not defined by place or by personal characteristics [10]. Online communities are not restricted by certain points in time but always open for action. Moreover, these communities are not

necessarily inhabited by individuals with a certain mind-set, age, or background, but often formed by members with a common interest or need [10]. Members in these globalised online communities may be politically, economically, and culturally indifferent on a local level, but still unite in communities in a quest to share information [11], and in our case to collectively conquer their disease.

Diabetes patients is to learn how to manage diabetes through several activities from clinical checks-up, blood sugar control, medical adherence, and lifestyle-changes. Generally, this knowledge may be learned through dialogue with healthcare professionals, caretakers, and disease-related materials, but these knowledge sources may also be combined with or followed up by knowledge-sharing in online communities [12]. As noted in a related study: *Virtual communities for diabetes health care play an important role in contributing to the overall effect of diabetes treatment worldwide*.

C. Patient adherence model

To emphasise the importance of online conversations for patients of diabetes, we now present a medical adherence model [13] which was originally introduced by Fisher et. al. to show how patient adherence is dependent on three vital elements: information, motivation, and behavioural skills.

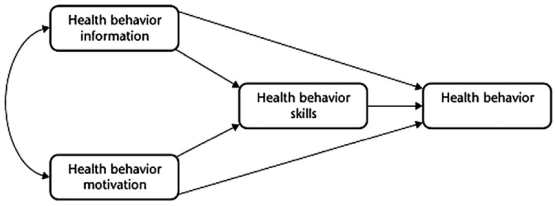

Figure 1. Information-Motivation-Behavioural Skills Model [13]

The model in fig. 1 depicts that adequate information related to a patient's personal treatment regimen is necessary for good adherence [14]. In addition, motivational factors, such as personal and social motivation to follow one's treatment regimen and the consequences of not adhering, are also significant. However, high motivation does not simply imply high information accuracy and vice versa. Behavioural skills are reflected in the actual performance of adherence. Behavioural skills encompass both the objective ability and the perceived efficacy of adhering to one's personal treatment regimen. Thus, the model has a direct relation between behavioural skills and adherence behaviour, while information and motivation play a vital part through behavioural skills, as seen by the arrows in fig. 1 [14]. We acknowledge that the model depicted here is a simplification of reality and that other factors such as environmental changes may also impact adherence. However, we have chosen to rely on this definition of the model as it represents the most crucial parts of medical adherence. We believe that accurate information and social support are fundamental in not only adherence, but in the general well-being of patients with diseases such as diabetes. The case study in [15] concludes that patients discussing their difficulties, conditions, and related problems in online communities become better at managing their disease and coping with related issues. This research builds on the idea that self-disclosure and social support is effective in treating and improving patient's self-efficacy.

We believe that the private and public healthcare system have a lot to learn from these online communities. By learning more about diabetes communities online, their topics, interests, and needs, we believe that health communication, information, and support directed towards the patients can be improved.

D. Health 2.0, support, and anonymity

Research has shown that patients use the internet more often than they communicate with their doctors about healthcare issues, and therfore they are increasingly interested in participating in online communities [16]. Researchers have termed these health communities as Health 2.0 [16] and this term builds on the related term Web 2.0, which is broadly defined as virtual relationships powered by social software. Likewise, Health 2.0 leverages social software in order to give patients new platforms to learn about their disease and get support from other patients with diverse experiences [16]. The research done by Greene et al. [16] shows that the conversations on the 15 most used diabetes pages on Facebook were concentrated on information sharing and social support, which validates our claim that online communities can be important in relation to the information-motivation-behavioural skills model.

Recent research [15] suggests that people are more likely to share sensitive information on Reddit compared to Facebook and these conversations on Reddit fulfilling information and social needs as part of the mental health discourse. It points out that Reddit enables one-time use of throw-away accounts, which facilitates anonymous conversations on sensitive topics, compared to other social media platforms such as Facebook, where user profiles usually are linked to personally identifiable information [15].

The related category of social platforms, known as bulletin boards and forums, has been one of the most popular sources for researching social communities online [17]. A reason might be that these forums afford users the ability to have a flexible identity and uphold anonymity via a self-chosen aliases. Such a pseudonym may be used to present the user's online identity when partaking in for instance disease related conversations, which may increase the propensity of community members to share and discuss issues, which they may find difficult to address in a real world situation. Moreover, forums are exclusive and topic specific, e.g. diabetes boards are not allowed to discuss cricket and vice versa. This in turn minimises the risk for condemnation and denunciation and gives the users a safe environment to seek information and support [18].

III. Research Methodology

Our full dataset consists of 39,425 texts collected from 42 online forums, including Reddit. From this dataset, a total of 3,309 randomly chosen texts were manually coded for 4 domain specific models (tab. I) and used as a training dataset. The data consists of peer-to-peer conversations about type 2 diabetes, originating from domains that are either dedicated to

diabetes or which have a health related focus. All the texts are in English and most of these originate from the US or UK, but also includes texts from India, Australia and other countries. The main filter was *diabetes*, with focus on type 2 diabetes and exclusion of type 1 diabetes. The dataset was collected on February 18th 2016 and the time-period is 27 months in total.

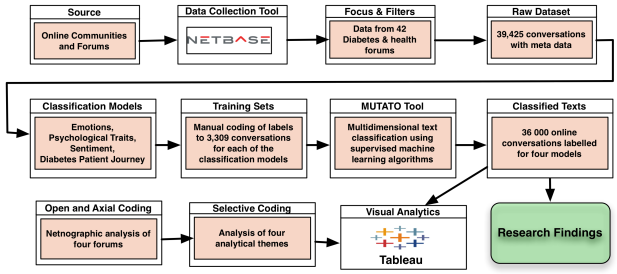

Figure 2. Data Analytic Process

As shown in Fig. 2, the overall process consists of three important steps: Data collection, data processing and data analysis. As part of data processing, we applied supervised machine learning technique, text classification with several algorithms (Fig. 4) to classify the online forum texts using our in-house and custom-build tool Multi-dimensional Text Analytics Tool (MUTATO). The overall architecture of MUTATO is shown in Fig. 3 and it can perform several text analysis tasks (both unsupervised and supervised) such as text mining, text classification and topic modeling using various open source machine learning libraries using Python as main programming language. As part of further data analysis we have used a qualitative approach netnographic analysis to get a better understanding about the general topics, interests, and challenges expressed in the data. Finally, we used visual analytics with help of Tableau to analyse our data and ultimately communicate our findings and insights.

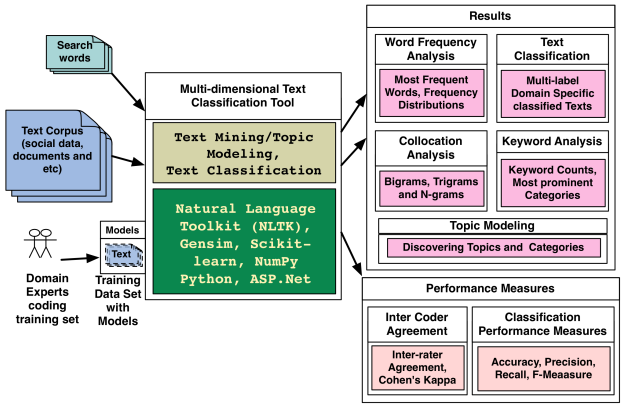

Figure 3. Multi-dimensional Text Analytics Tool (MUTATO) Architecture

A. Models for Text Classification

In order to analyse the online textual conversations, we have chosen four domain specific models (*Emotions*, *Sentiment*, *Personality Traits* and *Patient Journey*) based on their relevance and importance to our dataset and field of inquiry. As explained in Tab. I, the four models are: a model categorising the data into emotional stances, a model for sentiment, a model on personality traits, and a model categorising the data in respect to where a given author of the conversation might be in a diabetes patient journey, based on our interpretation of the authors' online expression. In order to get a deeper understanding of type 2 diabetes online discussions, the models are combined and results are juxtaposed. The chosen eight primary emotion labels (Tab. I) are adopted from [19] and it primarily consists of two main characteristics: they had to be identifiable at all phylogenetic levels and have adaptive significance in the individual's struggle for survival [19]. The big five personality traits (Tab. I) are adopted from [20]. A patient journey follows the patient from pre-diagnosis over treatment through to a viable cure [21]. We have chosen to

Label	Definition
Model 1: Emotions [19]	
Joy	Feeling of well-being, often also stated as happiness.
Sadness	Opposite of happiness or joy. Lowering of mood for a temporary period of time
Trust	Concerned with believing in something or a person.
Disgust	A feeling of revulsion or strong disapproval aroused by something unpleasant or offensive.
Fear	Fear is present when someone is trying to avoid some kind of pain or a threatening situation.
Anger	Intense emotional state that includes feelings such as irritation, provocation or even, at the extreme, rage.
Anticipation	Anticipation is a kind of expectation towards future. The expectation can be of a positive kind (feeling exited) or can be of fear or in extreme cases anxiety.
Surprise	Surprise is the result of experiencing something unexpected. Surprise is only momentarily.
Model 2: Sentiment	
Positive	Something is good and beneficial in a given context.
Neutral	Neither positive or negative.
Negative	Something is bad, hurtful or unwanted in a given context.
Model 3: Personality Traits	
Openness	Open to new ideas, experiences and is related to curiosity, adventure and imagination.
Conscientiousness	Someone who aims for achievements [20] and expresses a propensity to be thoughtful, thorough, in control, a preference for planning and structural living.
Extraversion	Extraversion is often opposed to Introversion [20] and the extravert is often the centre of attention, out-going, socially comfortable, energetic and likes to talk.
Agreeableness	Focused on establishing consensus to achieve social harmony. Such individuals often conform to social norms and are usually generous, trustworthy, optimistic, caring and emotionally supportive [20].
Neuroticism	This trait is linked to emotional instability, anxiety and depression [20]. Individuals labeled with neuroticism will be vulnerable and emotionally reactive.
Model 4: Patient Journey	
Undiagnosed	People without diabetes, patients with pre-diabetes or gestational diabetes and factual texts about diabetes
Relatives of diabetes patients	People discussing topics on behalf of family/friends diagnosed with diabetes or in the risk zone
Diagnosis	Patients diagnosed with diabetes by a health care professional.
Clinical Treatment	Everything related to medical treatment of diabetes. Clinical treatment, managing, adhering to treatment.
Alternative Treatment	Conversations related to alternative treatment (e.g. Ayurveda or home remedies).
Living with diabetes - Lifestyle, social & psychological	Everything related to managing social and psychological life related to diabetes. Topics may include how diabetes have changed the social lifestyle or affects the patient psychologically
Living with diabetes-nutrition	Includes discussions about diet, recipes and other questions related to nutrition.
Living with diabetes-exercise	Includes discussions and questions related to an active lifestyle

Table I
DESCRIPTION DOMAIN SPECIFIC MODELS FOR TEXT CLASSIFICATION

conduct our own Patient Journey for two reasons. Firstly, we were unable to find any theory or research that supports it and secondly, we wanted to make it more specific to the exact case and disease (diabetes) while being grounded in our dataset. This has led us to the Patient Journey domain specific model as shown in Tab. I. In order to achieve a good consistency of training sets, coding of all 3309 online texts is done by the first two authors individually and inter-rater agreement were calculated using Cohen's kappa coefficient [22] to make sure that agreement between the coders is fairly accurate.

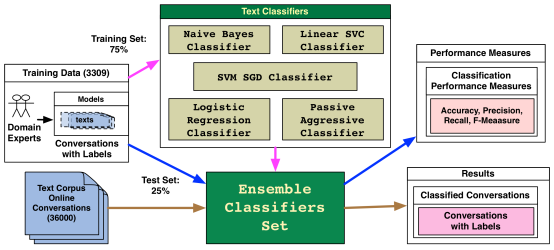

Figure 4. Text Classification Architecture

B. Text Classification Performance

Model 1: Emotions				
Classifiers	**Precision**	**Recall**	**F1-score**	**Accuarcy**
Multinomial NB	0.71	0.70	0.69	0.695
Linear SVC	0.75	0.75	0.74	0.748
Logistic Regression	0.68	0.58	0.51	0.576
Passive Aggressive	0.73	0.73	0.73	0.735
SVM SGD	0.66	0.61	0.58	0.613
Voted Accuracy	-	-	-	0.706
Model 2: Sentiment				
Classifiers	**Precision**	**Recall**	**F1-score**	**Accuarcy**
Multinomial NB	0.78	0.77	0.77	0.772
Linear SVC	0.81	0.81	0.81	0.806
Logistic Regression	0.73	0.7	0.67	0.698
Passive Aggressive	0.81	0.81	0.81	0.807
SVM SGD	0.73	0.62	0.53	0.621
Voted Accuracy	-	-	-	0.789
Model 3: Personality Traits				
Classifiers	**Precision**	**Recall**	**F1-score**	**Accuarcy**
Multinomial NB	0.67	0.66	0.66	0.661
Linear SVC	0.69	0.68	0.68	0.683
Logistic Regression	0.63	0.62	0.61	0.619
Passive Aggressive	0.69	0.69	0.69	0.691
SVM SGD	0.65	0.64	0.63	0.636
Voted Accuracy	-	-	-	0.675
Model 4: Patient Journey				
Classifiers	**Precision**	**Recall**	**F1-score**	**Accuarcy**
Multinomial NB	0.84	0.84	0.84	0.843
Linear SVC	0.87	0.87	0.87	0.872
Logistic Regression	0.81	0.79	0.78	0.794
Passive Aggressive	0.88	0.88	0.88	0.881
SVM SGD	0.8	0.79	0.77	0.786
Voted Accuracy	-	-	-	0.865

Table II
PERFORMANCE MEASURES OF THE CLASSIFIERS

As part of text classification, we have used ensemble learning method [23] by using 5 different supervised machine learning algorithms to build a set of text classifiers by using the voting method to predict most probable label for a given textual conversation from online communities. As shown in Fig. 4, we have used 5 different algorithms for text classification: 1) Multinomial Naïve Bayes classifier (Multinomial NB) [24] 2) Linear Support Vector Classifier (Linear SVC) [25] 3) Logistic Regression Max Entrophy classifier (Logistic Regression) [26] 4) Passive-Aggressive classifier (Passive Aggressive) [27] and 5) Support Vector Machines with stochastic gradient descent (SVM SGD) [28]. Text classifiers were built for these algorithms based on their respective implementation using scikit-learn [29] machine learning library in Python.

As shown in Fig. 4, classifiers were trained with using 75% of manually coded conversations as training set and the rest 25% were used as test set to predict the performance of the classifiers. The performance measures: Precision, Recall, F1-score and Accuracy of each classifier were computed and tabulated in Tab. II. All the classifiers more or less performed consistently with respect precision, recall, F1-score and the voted accuracies varied in between 0.675 to 0.865, which shows that the predictions of the classifiers was highly accurate. Among the classifier algorithms, both Linear SVC and Passive Aggressive performed better than the rest of the algorithms. In respect of the models, patient journey model received high voted accuracy which is 0.865. When once the classifiers are trained, the rest of the conversations (36000) were classified.

IV. RESULTS

Some of the important results are presented here.

1) High Degree of Trust in Online Forums: One of the first topics that emerged while coding was a strong confusion about diabetes. Many patients have unanswered questions and often do not know how the disease will affect them and how best to manage it. In this regard, we find that online communities play a vital role in knowledge-sharing and general discussions around disease-related information. Our analysis of conversations showed that people use the forum for information sharing activities. As we can see from Fig. 5, *conscientiousness* is the most prevalent personality trait among the users and when it combined with emotions it becomes pretty clear that the *trust* represent the largest share of conscientiousness posts. This result indicates that the users have high amount of *trust* in these online forums.

2) Support in the Digital Space: We have found that online communities not only exist to fulfil information needs, but also being supportive for patients seeking comfort and empathy. This is especially true for patients who have been diagnosed recently, as they often express a need for support and motivation. It creates uncertainty when research cannot provide a definite answer about what causes diabetes to the people. Some are blaming obesity, some are blaming the food habits, and others are suggesting that it is a genetic predisposition. Instead, people in online communities draw on each other's experiences. In our analysis, such empathy, social guidance, and support have been classified as *agreeableness* and as shown in Fig. 5 agreeableness is the major personality trait after the conscientiousness.

3) Perceptions of Health - Diet and Nutrition: One of the main concerns about living with type 2 diabetes is the constant struggle to control the body's blood glucose levels. To keep

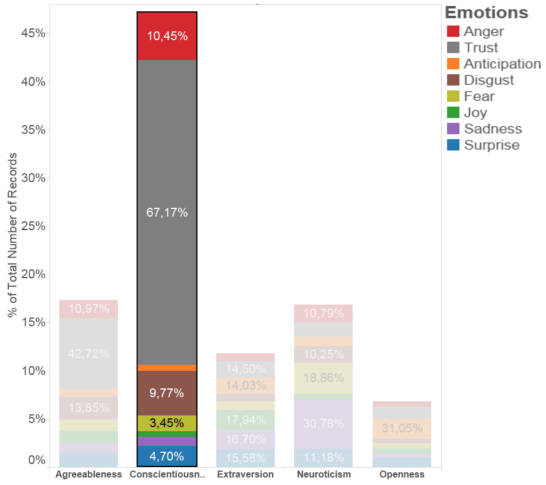

Figure 5. Data Analytic Process

the blood sugar low, many diabetes patients are prescribed metformin as part of their treatment regimen[1]. However, many patients advocate that diet and nutrition can either cure or at least slow down disease progression, and argue that monitoring one's blood glucose levels is a necessity. The focus on nutrition is highly prevalent in the discussions, as majority of posts classified in relation to the Patient Journey model are labelled with Nutrition. The amount of Nutrition posts exceeds the amount of any other category, which supports the idea of people either looking for dietary advice or helping others with their Nutrition and diets.

Furthermore, we wanted to see if there existed a high association between Exercise and Nutrition. We filtered out all the texts that were labelled Exercise with a probability of $\geq$ 0.51, and then calculated which other labels were associated to these. This calculation was done by counting how many times each label had a probability of $\geq$ 0.25, as this would indicate the second highest probability, when Exercise had received $\geq$ 0.51. The results confirmed our assumption that Nutrition had the highest association with the Exercise labels.

V. Discussion and Conclusion

In this paper, we have used supervised machine learning approach to analyse conversations of online communities using five different text classification algorithms. The conversations were analysed for the four domain specific models (Emotions, Sentiment, Personality Traits and Patient Journey) to identify the nature of these conversations and also important topics that were being discussed.

Information sharing and community knowledge: It has become clear from our analysis that forum members use online communities as a place where they can learn and share information on topics of interest. In addition, we have found that informational content is highly present in the forums that we have investigated.

[1]http://www.who.int/diabetes/global-report/en/

Void-filling communities: Through our analysis, we have pointed out that the healthcare system is often inadequate in delivering the necessary information and support to patients. The existence of this void seems to create greater information sharing between patients online, which also emphasises the importance of the learning that can be extracted from these communities.

Social cohesiveness: In our analysis, it became clear that many online forum members partake in supportive acts. A large share of members either ask directly for support or indicate that they need support. Accordingly, a large percentage of replies have a supportive answer to these requests. Our findings about support, socialisation, and information seeking, bring us to the information-motivation-behavioural skills model. Our findings reveal that online communities consist of supportive and motivational conversations along with a high degree of disease-specific information and management posts and also the amount of Trust exists in these forums is quite high.

The interest in diabetes among online forums seems to mirror the continuously growing impact that the diabetes has on individuals around the world. From our netnographic inquiries, it became clear that online communities with relation to diabetes have become ever more popular. The activity level on these forums is high and dynamic, as new members join and others leave. This is certainly true for diabetes.co.uk, which has almost doubled the number of members from 100,000 in 2014 to almost 200,000 by 2016. While the amount of forum members might be low compared to the most popular social media platforms, the topic specific nature of these forums makes them unique, as they become home to sensitive conversations about diabetes.

To answer our research question comprehensively, the conclusions have been divided into answering our three sub-questions. With our first sub-question, we wanted to investigate which main interests and challenges may be found from interpreting online patient conversations about diabetes. We have found that informational posts, often expressed as normative statements, are the most prevalent type of conversation surrounding type 2 diabetes. Moreover, individuals requesting information and support from other forum members are also highly prevalent in these forums. Moreover, our qualitative assessment through netnography revealed that a large share of members are in fact in need of information, which is leading to extensive knowledge sharing among peers in these forums. For the second sub question, we found that machine learning is quite helpful in analysing huge number of textual conversations and also the findings from the text classification tallies with the qualitative assessment made through netnography.

We have shown how online conversations can lead to important insights and these insights however add no value before they are acted upon. For this purpose, we have proposed a number of suggestions and two recommendations, which can assist in activating these insights and turn them into strategic initiatives. This was done to provide an answer to our third and final sub-question: In what way can insights from online conversations about type 2 diabetes lead to strate-

gic recommendations? From a social business and outside-in perspective, we acknowledge how data can be of strategic value for organisations to improve products and services. We have proposed that companies and organisations can integrate our learnings, insights, and models to take advantage of the knowledge pooled in online communities. By using our supervised machine models, pharmaceutical companies can perform repeated social media analytics on data from their target audiences and be responsive to patients' needs and struggles. The models may also be used to test brand opinion and company sentiment by applying them to new datasets. This may lead to tactics that are set out to improve the corporate image or brand. Companies could use their domain specific models for text classification to study the topics of conversations of their customers and stakeholders.

Our findings also conform that patients express the need for more and better qualified information. This demand for information may be due to a lack of informational supplies, which may be a result of insufficient, inaccurate, or at times even misleading information, given by some healthcare professionals which can be seen from the number of users expressing distrust on healthcare professionals and the healthcare system in these conversations. As a consequence, forums become important platforms for trusted knowledge sharing and articulation of diverse discourses around diabetes.

The need for support of diabetes patients was evident in our research. This was especially true with frustrated patients, who were recently diagnosed, and relatives of diabetes patients. These two categories also revealed a high degree of sadness, negativity, and hopelessness in their conversations. This was counteracted upon by other forum members who provided positive distanced support for these members. It shows us how forums can act as computer-mediated support platforms for patients in need and also how technology, through the concept of Health 2.0, empowers patients to manage their disease and possibly increase their quality of life.

ACKNOWLEDGMENT

The authors were partially supported by ReVus project, Big Data Analytics for Public Health funded by Copenhagen Health Innovation Fund. Any opinions, findings, interpretations, conclusions or recommendations expressed in this paper are those of its authors and do not represent the views of the funding agencies.

REFERENCES

[1] S. Fox and M. Duggan, "Health online 2013," *Washington, DC: Pew Internet & American Life Project*, 2013.

[2] A. Kotov, *Healthcare data analytics*. Chapman and Hall/CRC, 2015, ch. Social Media Analytics for Healthcare., pp. 309–340.

[3] N. Straton, K. Hansen, R. R. Mukkamala, A. Hussain, T.-M. Gronli, H. Langberg, and R. Vatrapu, "Big social data analytics for public health: Facebook engagement and performance," in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2016, pp. 1–6.

[4] D. C. K. af Rosenborg, I. Buhl-Andersen, L. B. Nilsson, M. P. Rebild, R. R. Mukkamala, A. Hussain, and R. Vatrapu, "Buzz vs. sales: Big social data analytics of style icon campaigns and fashion designer collaborations on h&m's facebook page," in *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*, 2017.

[5] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Z. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, Jan 2017.

[6] N. Straton, R. R. Mukkamala, and R. Vatrapu., "Big social data analytics for public health: Predicting facebook post performance using artificial neural networks and deep learning." in *In proceedings of IEEE International Congress on Big Data (IEEEBigdata-2017)*, 2017.

[7] S. P. Collins, "The youngest case of type 2 diabetes ever recorded highlights an 'emerging epidemic'," https://thinkprogress.org/the-youngest-case-of-type-2-diabetes-ever-recorded-highlights-an-emerging-epidemic-2015.

[8] G. Day and C. Moorman, *Strategy from the outside in: Profiting from customer value*. McGraw Hill Professional, 2010.

[9] G. J. Johnson and P. J. Ambrose, "Neo-tribes: The power and potential of online communities in health care," *Communications of the ACM*, vol. 49, no. 1, pp. 107–113, 2006.

[10] E. Wenger, N. White, and J. D. Smith, *Digital habitats: Stewarding technology for communities*. CPsquare, 2009.

[11] R. Robertson *et al.*, "Glocalization: Time-space and homogeneity-heterogeneity," *Global modernities*, vol. 2, pp. 25–45, 1995.

[12] S. Ilioudi, A. A. Lazakidou, N. Glezakos, and M. Tsironi, "Health-related virtual communities and social networking services," in *Virtual Communities, Social Networks and Collaboration*. Springer, 2012, pp. 1–13.

[13] W. A. Fisher, J. D. Fisher, and J. Harman, "The information-motivation-behavioral skills model: A general social psychological approach to understanding and promoting health behavior," *Social psychological foundations of health and illness*, pp. 82–106, 2003.

[14] K. Amico, J. Toro-Alfonso, and J. D. Fisher, "An empirical test of the information, motivation and behavioral skills model of antiretroviral therapy adherence," *AIDS care*, vol. 17, no. 6, pp. 661–673, 2005.

[15] M. De Choudhury, "Opportunities of social media in health and well-being," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 21, no. 2, pp. 23–27, 2014.

[16] J. A. Greene, N. K. Choudhry, E. Kilabuk, and W. H. Shrank, "Online social networking by patients with diabetes: a qualitative evaluation of communication with facebook," *Journal of general internal medicine*, vol. 26, no. 3, pp. 287–292, 2011.

[17] R. V. Kozinets, *Netnography: Doing ethnographic research online*. Sage publications, 2010.

[18] R. Langer and S. C. Beckman, "Sensitive research topics: netnography revisited," *Qualitative Market Research: An International Journal*, vol. 8, no. 2, pp. 189–203, 2005.

[19] R. Plutchik, *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association, 2003.

[20] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual review of psychology*, vol. 41, no. 1, pp. 417–440, 1990.

[21] Dartmouth-Hitchcock, "A typical patient's journey: Diabetes," http://www.dartmouth-hitchcock.org/diabetes/a_typical_patients_journey_diabetes.html, 2017.

[22] K. Krippendorff, "Content analysis," in *International Encyclopedia of Communication*. Oxford University Press, 1989.

[23] T. G. Dietterich, "Ensemble learning," *The handbook of brain theory and neural networks*, vol. 2, pp. 110–125, 2002.

[24] 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. Madison, WI, 1998, pp. 41–48.

[25] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[26] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in neural information processing systems*, 2002, pp. 841–848.

[27] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, no. Mar, pp. 551–585, 2006.

[28] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," *Mathematical programming*, vol. 127, no. 1, pp. 3–30, 2011.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2078195