CHIS@FIRE: Overview of the Shared Task on Consumer Health Information Search

Manjira Sinha, Sandya Mannarswamy, Shourya Roy
Xerox Research Center India
Bengaluru, India
(manjira.sinha, sandya.mannarswamy, shourya.roy)@xerox.com

ABSTRACT

People are increasingly turning to the World Wide Web to find answers for their health and lifestyle queries, While search engines are effective in answering direct factual questions such as 'What are the symptoms of a disease X?', they are not so effective in addressing complex consumer health queries, which do not have a single definitive answer, such as 'Is treatment X effective for disease Y?'. Instead, the users are presented with a vast number of search results with often contradictory perspectives and no definitive conclusion. The term "Consumer Health Information Search" (CHIS) is used to denote such information retrieval search tasks, for which there is "No Single Best Correct Answer". The proposed **CHIS track** aims to investigate complex health information search in scenarios where users search for health information with more than just a single correct answer, and look for multiple perspectives from diverse sources both from medical research and from real world patient narratives.

Keywords

information retrieval for clinical texts, multi-perspective health data mining

1. INTRODUCTION

World Wide Web is increasingly being used by consumers as an aid for health decision making and for self-management of chronic illnesses as evidenced by the fact that one in every 20 searches on Google [5] is about health. Information access mechanisms for factual health information retrieval have matured considerably, with search engines providing Fact checked Health Knowledge Graph search results to factual health queries. While the direct informational needs of the *Online Health Information Seekers* regarding well established disease symptoms and remedies are well met by search engines [5], general search engines do not provide definitive answers for addressing complex consumer health queries which have multiple different points of view/perspectives associated with them.

It is pretty straightforward to get an answer to the query "what are the symptoms of Diabetes" from the search engines. However retrieval of relevant multiple perspectives for complex health search queries which do not have a single definitive answer still remains elusive with most of the general purpose search engines. For example, a user health query such as "can metabolic therapy cure brain cancer" causes considerable frustration for the searcher as he needs to wade through hundreds of search results to obtain a balanced view of the diverse perspectives/points of view available, both for and against the hypothesis posed in the search query. Subjective health related queries such as *'does treatment X effective for disease Y?'* or *'can X cause disease Y'* do not have a single definitive answer on the web due to the multiple supporting/opposing perspectives available on the web related to them, instead multiple perspectives (which very often are contradictory in nature) are available on the web regarding the queried information. The presence of multiple perspectives with different grades of supporting evidence (which is dynamically changing over time due to the arrival of new research and practice evidence) makes it all the more challenging for a lay searcher. Figure 1 depicts the precise scenario. In our *Consumer Health Information search* **CHIS** shared task track on FIRE [1], we have attempted to encourage the development of innovative computational models to statistically represent the multiple-perspective around a general health search query and therefore, assist the self-searcher with better and meaningful information insights..

Can metabolic therapy cure brain cancer?	
Support	**Oppose**
Glucose is the primary fuel to cancer cells, hence low carb ketogenic diets can be effective in reducing growth of cancer cells in brain	Cancer is a genetic disease, hence cannot be impacted by any metabolic therapy
There are two specific patient narratives reported in literature, discussing specific ketogenic diets, which had led to remission of brain cancer	A Cochrane review found no significant evidence for the effectiveness of metabolic therapy in curing brain cancer.

Figure 1: Contradicting search results for clinical query

2. BACKGROUND

At present times, there has been considerable interest in the field of stance classification and stance modelling. Stance classification has been applied to different debate settings such as congressional debates [15, 18, 3], company internal debates [9, 10, 1] and online public forums on social and political topics [13, 14, 17, 4, 16, 2]. Recently there has been work on stance classification of argumentative political essays [6], online news articles [7] and online news comments [12].

Unlike many of the earlier research settings which have analyzed posts on public debate topics, multi-perspective con-

[1]https://sites.google.com/site/multiperspectivehealthqa/

sumer health information is not typically characterized by strong emotion/opinion bearing language, nor does it have strongly delineated supporting/opposing topic words. They typically contains domain specific technical terms and sparse in emotional/affective words and is typically factual in nature. A closely related work [19] discussed the information seeking behaviour on MMR vaccine on internet search engines and developed an automated way to score Internet search queries and web pages as to the likelihood of the searcher deciding to vaccinate. Also while socio-political debate stances can often be delineated by well demarcated topic words (for instance, pro-abortion stance is often characterized by the topical words *'right to choose'*, whereas anti-abortion is characterized by the topical words *'pro-life'*), health related texts do not typically contain stance delineating topic words since the same proposition can be used for supporting or opposing a given health query, depending on the supporting research evidence. For instance, consider the following example sentences retrieved in response to the query ***Sun exposure causes skin cancer***:

- S1: Many studies have found that skin cancer rates are increasing in indoor workers.
- S2: very few studies have demonstrated that skin cancer rates are increasing in indoor workers.

Both sentences contain the topical phrase *skin cancer rates in indoor workers* with sentence S1 providing evidence in support of it, whereas sentence S2 providing evidence opposing it. This illustrates the difficult of identifying stance delineating topic words in health related text.

The technical language of the information in these queries is also another factor which makes the stance classification complex. Given an example sentence *E-cigarettes contain di-acetyl which has been associated with popcorn lung syndrome* for a sample query *E-cigarettes are safer than normal cigarattes*, it is not evident at first glance, whether this sentence is supportive/opposing the query. This makes the task more challenging, compared to general domain stance classification.

3. TASK DESCRIPTION

Given a CHIS query, and a document/set of documents associated with that query, the task is to classify the sentences in the document as relevant to the query or not. The relevant sentences are those from that document, which are useful in providing the answer to the query. These relevant sentences need to be further classified as supporting the claim made in the query, or opposing the claim made in the query.

Example query: Does daily aspirin therapy prevent heart attack?

S1: "Many medical experts recommend daily aspirin therapy for preventing heart attacks in people of age fifty and above." [**affirmative/Support**]

S2: "While aspirin has some role in preventing blood clots, daily aspirin therapy is not for everyone as a primary heart attack prevention method." [**disagreement/Oppose**]

3.1 Detailed Task Description

There are two sets of tasks:

1. **TASK A:** Given a CHIS query, and a document/set of documents associated with that query, the task is to classify the sentences in the document as relevant to the query or not. The relevant sentences are those from that document, which are useful in providing the answer to the query.

2. **TASK B:** These relevant sentences need to be further classified as supporting the claim made in the query, or opposing the claim made in the query.

Example :

- *Query-* "Are e-cigarettes safer than normal cigarettes?"

- *Retrieved sentence S1-* "Because some research has suggested that the levels of most toxicants in vapor are lower than the levels in smoke, e-cigarettes have been deemed to be safer than regular cigarettes". **A)Relevant, B) Support**

- *Retrieved sentence S2-* "David Peyton, a chemistry professor at Portland State University who helped conduct the research, says that the type of formaldehyde generated by e-cigarettes could increase the likelihood it would get deposited in the lung, leading to lung cancer." **A)Relevant, B) Oppose**

- *Retrieved sentence S2-* "Harvey Simon, MD, Harvard Health Editor, expressed concern that the nicotine amounts in e-cigarettes can vary significantly." **A)Irrelevant, B) Neutral**

Our task have 5 consumer health queries, Figure 2 and figure 3 below presents the comprehensive statistics of the CHIS queries used in our task released as training and test respectively.

Query	Support	Oppose	Neutral
E-cigarettes are safer than normal cigarettes (EC)	93	165	155
Sun exposure leads to skin cancer (SC)	105	78	158
Vitamin C prevents common cold (VC)	111	68	99
Women should take HRT post menopause (HR)	42	136	68
MMR vaccine can cause autism (MR)	72	94	93

Figure 2: Statistics of Queries in Training Data.

query	total	support	oppose	neutral
ecig	64	21	15	28
Sun exposure	88	34	3	51
Vitamin C	74	19	21	34
HRT	72	31	20	21
MMR	58	17	32	9

Figure 3: Statistics of Queries in Test Data.

4. TASK PARTICIPANTS AND RESULTS

A total of 9 teams participated in task and 9 submissions are obtained against Task A and 8 submissions are obtained against Task B. Details of the participating teams are shown in figure 4 below.

Team Name	Team Members	Institute
Amrita_CEN	Renmiya devi G, Veena P V , Anand Kumar M and Soman K P	Amrita University, Coimbatore
Amrita_Fire_CEN	Barathi Ganesh HB, Dr. Anand Kumar M, Dr. Soman KP	Tata Consultancy Services, Kochi; Amrita Vishwa Vidyapeetham, Coimbatore
JNTUH	S Suresh Kumar, L Naveen	JNTU Hyderabad; BVRIT, Hyderabad
JU_KS_Group	Kamal Sarkar, Indra Banerjee, Debanjan Das, Mamta Kumari ,Prasenjit Biswas	Jadavpur University, Kolkata
Fermi	Vijayasaradhi, Subba Reddy	International Institute of Information Technology-Hyderabad (IIIT- H)
Techie-challengers	Raksha Jalan , Nikhil Priyatam Pattisapu, Vasudeva Varma	International Institute of Information Technology-Hyderabad (IIIT- H)
SSN_NLP	Dr. D. Thenmozhi, Dr. P. Mirunalini and Dr. C Aravindan	SSN College of Engineering, Chennai
Individual	Jainisha Shankhavara	DAIICT, Gandhinagar
Individual	Hua Yang, Teresa Gonçalves	University of Évora, Portugal

Figure 4: Team details

4.1 Performance of Teams in Task A

Figure 5 below presents the team performance statistics for Task A, i.e., where a retrieved instance has to be classified according to whether it is relevant or irrelevant to a specific query[2].

TASK A		Accuracy (%)								
Query		Amrita_fire_CEN	JNTUH	Fermi	JU_KS_Group	Techie challengers	SSN_NLP	Amrita_cen	Hua Yang	Jainisha Sankhavara
Does Sun exposure cause skin cancer?		54.55	62.50	78.40	48.86	68.18	79.55	48.86	53.41	52.27
Can MMR vaccine lead to autism in children?		87.93	56.90	79.31	89.66	87.93	81.03	88.89	84.48	87.93
Should women take Hormone Replacement Therapy (HRT) post menopause?		70.83	38.89	88.88	93.06	75.00	87.50	75.86	90.28	91.67
Are E-cigerettes safer than normal cigerettes?		71.88	57.81	65.62	71.88	71.88	64.06	76.56	46.88	54.69
Does Vitamin C prevent common cold?		55.41	58.11	72.97	63.51	62.16	78.38	60.81	71.62	64.86
Overall Mean Accuracy		68.12	54.84	77.04	73.39	73.03	78.10	70.20	69.33	70.28

Figure 5: Final Result Table for Task A

As can be observed in Task A, team SSN_NLP and team Fermi have secured the top scoring positions with accuracy 78.10% and 77.04% respectively. SSN_NLP has proposed a decision tree model based on sophisticated text features including part-of speech. They have used a chi-square feature selection to extract the informative features and reduce the number of spurious features and demonstrated that such a feature selection approach can offer a significant gain. Team Fermi have used a deep neural network architecture with Rectified-linear (ReLu) and Sigmoid activation over bag-of-phrase features.

Team JU_KS_group and Techie-challengers have secured the second position jointly with a closed call of 73.39% and 73.03% accuracy respectively. JU_KS_group has implemented a support vector machine with polynomial kernel to classify the data. They have curated informative text features such as part-of-speech matching, neighborhood matching to represent the input data. Techie-challengers has proposed a naive-bayes classifier on doc2vec [8] and tf-idf based ensemble representation of the data.

With accuracy 70.20% and 70.28%, team Amrita_Cen and individual participant Jainisha Shankhavara have ranked third jointly. Team Amrita_Cen has used a support-vector-machine classifier on top of input feature representation obtained by word-embedding and keyword generation techniques. Jainisha has proposed classification model based on BM-25 [11] ranking function and tf-idf based input representation.

Hua Yang has approached the task from the perspective of improving understandability in consumer health related searches and their information retrieval based query expansion module has provided a 69.33% accuracy.

Team Amrita_Fire_Cen has used a random forest classifier on distributional semantic representation of the input and obtained 68.12% accuracy. They have used the non-negative matrix factorization technique for obtaining the distributed representation.

Team JNUTH model uses a aggregate over a range of similarity measures to obtain the relevance-irrelevance decision for a data input. They have obtained 54.84% accuracy.

[2]This is the final updated result table, the individual team working notes may not contain the latest updated version due to some late changes

4.2 Performance of Teams in Task B

Figure 6 below demonstrates the team performances for Task B.

TASK B		Accuracy (%)								
Query		Amrita_fire_CEN	JNTUH	Fermi	JU_KS_Group	Techie challengers	SSN_NLP	Amrita_cen	Hua Yang	Jainisha Sankhavara
Does Sun exposure cause skin cancer?		56.82	64.77	73.80	44.32	62.50	0.00	23.86	46.59	37.50
Can MMR vaccine lead to autism in children?		32.76	65.52	44.82	32.76	68.97	0.00	34.72	63.79	46.55
Should women take Hormone Replacement Therapy (HRT) post menopause?		26.39	48.61	54.16	22.22	37.50	0.00	43.10	48.61	27.78
Are E-cigerettes safer than normal cigerettes?		37.50	67.19	51.56	29.69	60.94	0.00	39.06	60.94	46.88
Does Vitamin C prevent common cold?		39.19	31.08	50.00	39.19	32.43	0.00	32.43	50.00	31.08
Overall Mean Accuracy		38.53	55.43	54.87	33.64	52.47	0.00	34.64	53.99	37.96

Figure 6: Final Result Table for Task B

In task B, team JNUTH has jointly secured the first position with team Fermi. JNUTH has used a C-support vector machine classifier with radial basis kernel. They have used tf-idf for input representation followed by a max-feature sorting. Their model has obtained 55.43% accuracy. Team Fermi has used a deep neural network architecture and a bag-of-phrase representation to achieve 54.87% accuracy.

With a score of 53.99% Hua Yang have secured the second rank. His model uses a naive-Bayes classifier and tf-idf representation. Team Techie-challengers also used a naive-Bayes classifier, but on doc2vec input representation to obtain 52.47% accuracy. Therefore, they hold the third rank.

Team Amrita_Fire_Cen has used a random forest classifier on distributional semantic representation of the input and obtained 38.53% accuracy. Individual participant Jainisha Sankhavara has developed a model based on BM-2 ranking function to obtain overall accuracy 37.96%.

Team Amrita_Cen has modeled using a support vector machine classifier with input feature representation obtained by word-embedding and keyword generation techniques to obtain 34.64% accuracy. Team JU_KS_group has modeled the task as sentiment classification problem and their innovative feature set consists of positive, negative and neutral polarity words along with information from Task A. They have achieved an overall accuracy of 33.64%.

5. CONCLUSION

We thank all the participants for expressing interest in our track. It has been a great experience to witness the innovative models and techniques proposed by different teams. The CHIS task was surely a challenging one with little presiding literature and yet, as can be observed from the previous section, in both the tasks there are closed calls in terms of the performances of different teams.

We also express our sincere gratitude to the organizing and program committee of **Forum for Information Retrieval Evaluation (FIRE), 2016**, especially Mr. Parth Mehta, for providing us with the opportunity to hold the shared task and to connect with the enthusiast researchers across India and abroad who share the same interest.

In future, we are looking forward to work again with such expert groups to come up with novel solutions to more challenging health-care data analytic problems.

6. REFERENCES

[1] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 529–535, New York, NY, USA, 2003. ACM.

[2] P. Anand, M. Walker, R. Abbott, J. E. F. Tree, R. Bowmani, and M. Minor. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 1–9, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[3] A. Balahur, Z. Kozareva, and A. Montoyo. Determining the polarity and source of opinions expressed in political debates. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '09, pages 468–480, Berlin, Heidelberg, 2009. Springer-Verlag.

[4] O. Biran and O. Rambow. Identifying justifications in written dialogs,. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, 2011.

[5] O. G. Blog. Google health information knowledge graph, 2015.

[6] A. Faulkner. Automated classification of stance in student essays: An approach using stance target information and thewikipedia link-based measure. In *Proceedings of the Twenty-Seventh International Artificial Intelligence Research Conference*, 2012.

[7] W. Ferreira and A. Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics.*, 2016.

[8] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[9] T. Mullen and R. Malouf. Taking sides: User classification for informal online political discourse. *Internet Research*, 18:177–190, 2008.

[10] A. Murakami and R. Raymond. Support or oppose?: Classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 869–875, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[11] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49. ACM, 2004.

[12] P. Sobhani, D. Inkpen, and S. Matwin. From argumentation mining to stance classification. *NAACL HLT 2015*, 2015.

[13] S. Somasundaran and J. Wiebe. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 226–234, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[14] S. Somasundaran and J. Wiebe. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 116–124, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[15] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 327–335, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[16] M. A. Walker, P. Anand, R. Abbott, and R. Grant. Stance classification using dialogic properties of persuasion. NAACL HLT '12, pages 592–596, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[17] Y.-C. Wang and C. P. Rosé. Making conversational structure explicit: Identification of initiation-response pairs within online discussions. In *2010 North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 673–676, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[18] A. Yessenalina, Y. Yue, and C. Cardie. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1046–1056, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[19] E. Yom-Tov and L. Fernández-Luque. Information is in the eye of the beholder: Seeking information on the MMR vaccine through an internet search engine. In *AMIA 2014, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 15-19, 2014*, 2014.