1 Introduction

This project was set within the framework of a collaboration with the ALIAE startup where the aim is to develop a chatbot to collect information from clinical patients. The main idea is to replace strictly defined surveys by a more natural conversation in order to let users express themselves freely so that more information could be collected. First, we should be able to analyse the user utterances (does the user speaks about pain, physical activity etc. ?) and detect at least one most probable intent of user message so that chatbot could choose the right strategy for fulfilling information.

A small manually annotated dataset (roughly 100 sentences for each category) was available. So the purpose of the project is to find word embeddings that will allow to expand it by external data using the existing dataset with labelled intents.

We have tried models for sentence representation by word and the whole context in order to use semi-supervised learning by gradually increasing train data by labelling external sources.

2 Literature review

James Ferguson et al faced a similar problem with a small labelled dataset. In their article they describe the approach of automatically expanding dataset and improving the performance of baseline models, that was event trigger identification system in their case.

First, they trained a baseline classifier on available data. Then they identified clusters of additional data to obtain grouped paraphrases inspired by the NewsSpike idea introduced in Zhang et al. (2015). After they labelled clusters with baseline model trained fully-supervised. Combining new labelled data and original one they retrained the event extractor. (Semi-Supervised Event Extraction with Paraphrase Clusters: http://aclweb.org/anthology/N18-2058)

3 Methodology

sentence representation, clustering , classification

3.1 Sentence representation models

3.1.1 By Word representation

The most popular word-basis techniques to represent sentences are LDA (Latent Dirichlet Allocation) and Word2Vector models. We used their implementation from gensim library in order to build models based on the initial labelled dataset.

Preprocessed tokenized sentences were used to create dictionary and were passed to W2V model as it is and in form of Bag-Of-Word to LDA model.

For selecting the best model we changed space dimension (topics for LDA and features for W2W). Preferred number of features for W2V model is 200. It's the start point for our model.

The good point of using this models is that they can be used for both word and sentence representation since one word can be treated as a small sentence.

3.1.2 By Sentence representation

In order to use context of words more PositionwiseFeedForward and BiLSTM neural networks were implemenent on Pytorch.

3.2 Clustering models

The main idea behind using clustering models for evaluating sentence representations is to detect whatever sentences with the same intent populate the same clusters.

For these purposes following clustering models were used:

- K-Means (KM)
- AgglomerativeClustering(ward) (AG)
- GaussianMixture (GM)

The number of clusters was set to the number of intents (20) that allows using not only purity and Silhouette coefficients to evaluate them but also homogeneity and completeness to get the idea how resulting groups correlate with initial intents. Also, for each model, we plotted a confusion matrix with a background gradient to visualize how well clusters separate different intents.

3.3 Classification models

SVC on labelled data

Apply on unlabelled data

4 Datasets

4.1 Dataset with intents

The initial dataset is created manually covering 20 main possible users intents. Each sentence represents one intent. For, example,

```
{'text': 'After sleep, for 2-3 hours,
I am better and then start feeling tired again',
 'intent': 'sleep',
}
```

Distribution of labels is show on fig. 1

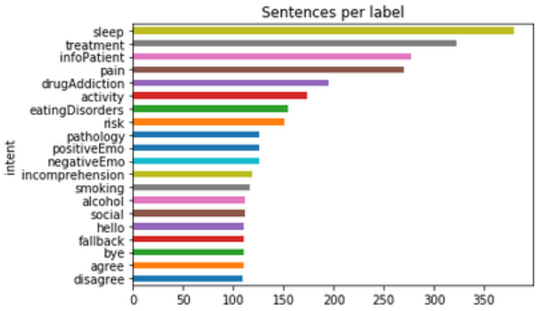


Figure 1: Label distribution in dataset

'min': 0, 'max': 12, 'mean': 3.438, 'std': 2.16367

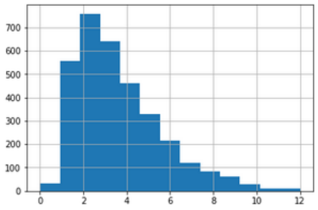


Figure 2: Sentence dist

Category of board — of posts	
digestive-disorders	5064
addiction-recovery	3644
sleep-disorders	1748
smoking-cessation	937
eating-disorder-recovery	762
chronic-pain	735
chronic-fatigue	662
stress	415
family-friends-addicts-alcoholics	312
pain-management	25

Table 1: Selected categories

Total number of sentences is 3305. Found 1882 words after filtering through stopwords list.

4.2 Forum data

To improve initial classifier we needed to extend dataset by much more data and more naturally constracted one. No test set is available was available at the beginning of the project. To solve this problem, we opted to created our own test set following Zhang et al (2015) and Sondhi et al (2010) example.

Finally, 272552 unique posts from 238 categories. Because our intents are not that precise we should select particular boards with relevant information to be able gradually expand their context. Result is in table **??**

4.3 Preprocessing routine

Some proprocessing routine should be covered. This routine includes sentence tokenization, word tokenization, stop words removing and lemmatization.

Most popular and common tools for nlp are NLTK and SpaCy but they differ in behaviour.

For example, in word tokenization they give different results that can influence not only simple statistics but meaning too. Some example of different tokenization can be seen in table 1. Though concating words to ones like 'flulike' or '35mg' or 'longterm' sometimes gives more robust and concrete meaning in case of big dataset, in our case it seems better to stay with spacy way of tokenization in order to have smaller and more simple vocabualary.

For stopwords removing there were three options: nltk, spacy and the longest one. The last option was rejected due to containing words like 'want', 'stop', 'successfully' etc. that can be useful for detecting basic intents like positive or negative emotion, social. Finally nltk one was selecting because of containing shorts like 'm' from 'am', 've' from 'have'. Final dictionary contained 1882 words. Also all numbers were changed to num. Chart (8) looks fine.

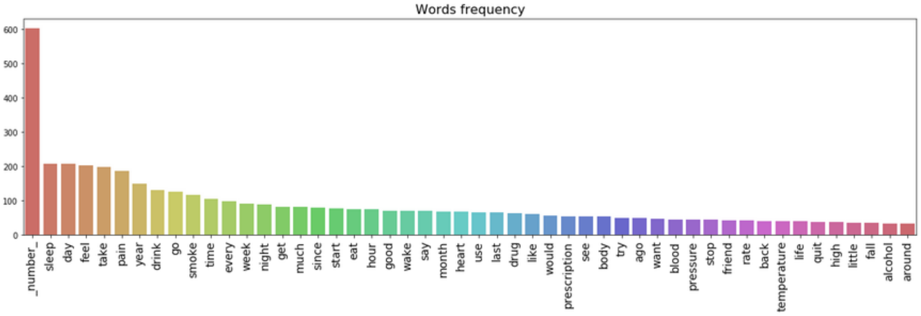


Figure 3: Words frequency

Length ¡ 30.

Next problem is empty sentences. But they don't change the dataset much.

intent fallback 12 disagree 11 hello 4 agree 3 incomprehension 2 positiveEmo 1

array(['what?', 'same that again', 'I am up', 'same', 'can you', 'do that', 'do this', 'Will do', 'no', 'no i will not', "No i don't", 'no', 'no i will not', "No i don't", "What's up?", "What's up?", 'he', 'a', 'd', 'i', 'm', 'o', 's', 't', 'y', 'I will', 'Will do', 'No', 'no', 'No it is not', "No I don't", 'No', "I'm here"], dtype=object)

HealthBoards is a medical forum web portal that allows patients to discuss their ailments. We scraped 272553 unique posts contained in each category. Finally, the corpus consists of N sentences. Table 4 shows the dataset statistics.

NLTK	SpaCy
[’i’, ’wouldnt’, ’go’, ’to’, ’sleep’, ’until’, ’like’, ’5’, ’6’, ’or’, ’8am’]	[’i’, ’would’, ’nt’, ’go’, ’to’, ’sleep’, ’until’, ’like’, ’5’, ’6’, ’or’, ’8’, ’am’]
[’that’, ’is’, ’totally’, ’wrongheaded’]	[’that’, ’is’, ’totally’, ’wrong’, ’headed’]
[’i’, ’am’, ’in’, ’the’, ’process’, ’of’, ’tapering’, ’from’, ’suboxone’, ’longterm’, ’use’]	[’i’, ’am’, ’in’, ’the’, ’process’, ’of’, ’tapering’, ’from’, ’suboxone’, ’long’, ’term’, ’use’]
[’i’, ’had’, ’an’, ’onandoff’, ’opiateopioid’, ’habit’, ’from’, ’about’, ’2010’]	[’i’, ’had’, ’an’, ’on’, ’and’, ’off’, ’opiate’, ’opioid’, ’habit’, ’from’, ’about’, ’2010’]
[’i’, ’have’, ’flulike’, ’pathologysymptom’]	[’i’, ’have’, ’flu’, ’like’, ’pathologysymptom’]
[’i’, ’have’, ’exerciseinduced’, ’insomnia’]	[’i’, ’have’, ’exercise’, ’induced’, ’insomnia’]
[’i’, ’m’, ’supposed’, ’to’, ’take’, ’6’, ’35mg’, ’tablets’, ’a’, ’day’, ’but’, ’i’, ’have’, ’taken’, ’20’, ’today’]	[’i’, ’m’, ’supposed’, ’to’, ’take’, ’6’, ’35’, ’mg’, ’tablets’, ’a’, ’day’, ’but’, ’i’, ’have’, ’taken’, ’20’, ’today’]

Table 2: Tokenization comparision

’min’: 0, ’max’: 2027, ’mean’: 10.932588521491454, ’std’: 8.964716092053479 ’min’: 3.2580388329566468, ’max’: 35.23600634007455, ’mean’: 11.224462266261876, ’std’: 6.963225985041148

Data should be divided in subsets by increasing sentence length because of the difference in mean values for both datasets.

5 Experiment

5.1 Clustering

5.1.1 Evoluation metrics

For every model following metrics and their average were calculated: purity score, Silhouette Coefficient, homogeneity and completeness scores. Later for each parameter average among clustering models calculated for each score in order to get both table and plot.

Also for each model confusion matrix created between cluster labels and initial intents.

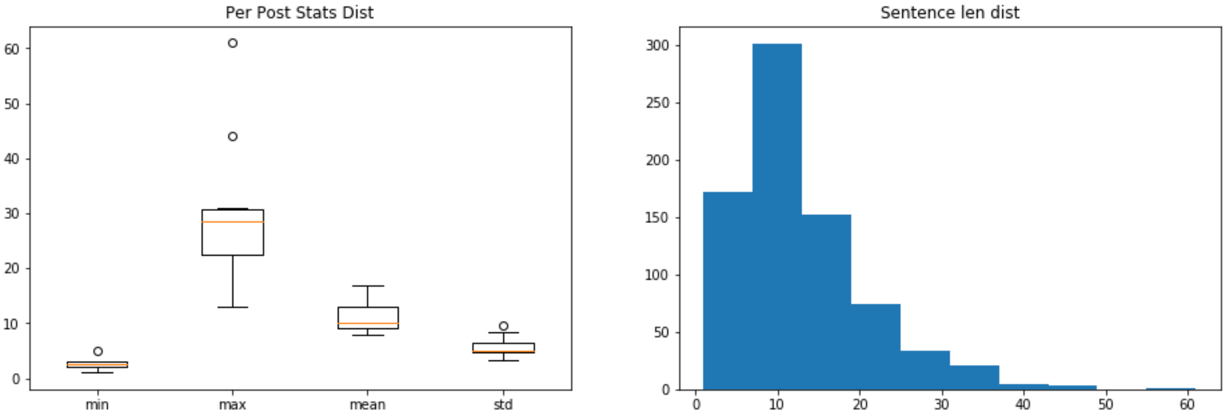


Figure 4: Text stat

5.1.2 W2V model

Comparision table

WE	Cluster	purity	Silhouette	homogeneity	complete	Clf CV score
Defenders	KM AC GM					
M	KM AC GM					
Forward	FW					
S	KM AC GM					

features num	purity	silhouette	homogeneity	completeness	AVG
10	0.192032	0.0774036	0.108761	0.127672	0.126467
20	0.186283	0.061294	0.108755	0.127291	0.120906
30	0.184065	0.0630405	0.108515	0.128856	0.121119
50	0.17761	0.0662788	0.107623	0.128358	0.119967
100	0.184266	0.0653807	0.109944	0.144074	0.125916
150	0.178417	0.0566761	0.105809	0.138289	0.119798
200	0.179929	0.0496567	0.102913	0.139742	0.11806
300	0.176803	0.0505959	0.105242	0.150409	0.120762
400	0.169844	0.072453	0.0984974	0.144132	0.121231
500	0.169541	0.0478535	0.0982757	0.15098	0.116663
550	0.172264	0.0620061	0.100714	0.153673	0.122164
600	0.170045	0.0498559	0.0988441	0.157109	0.118964

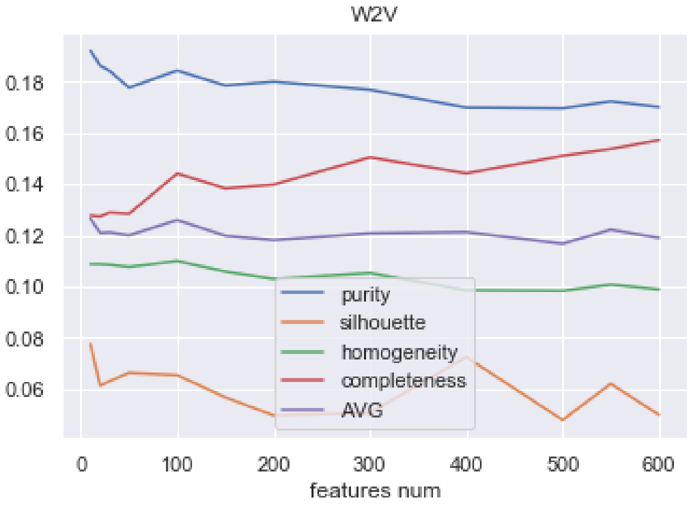
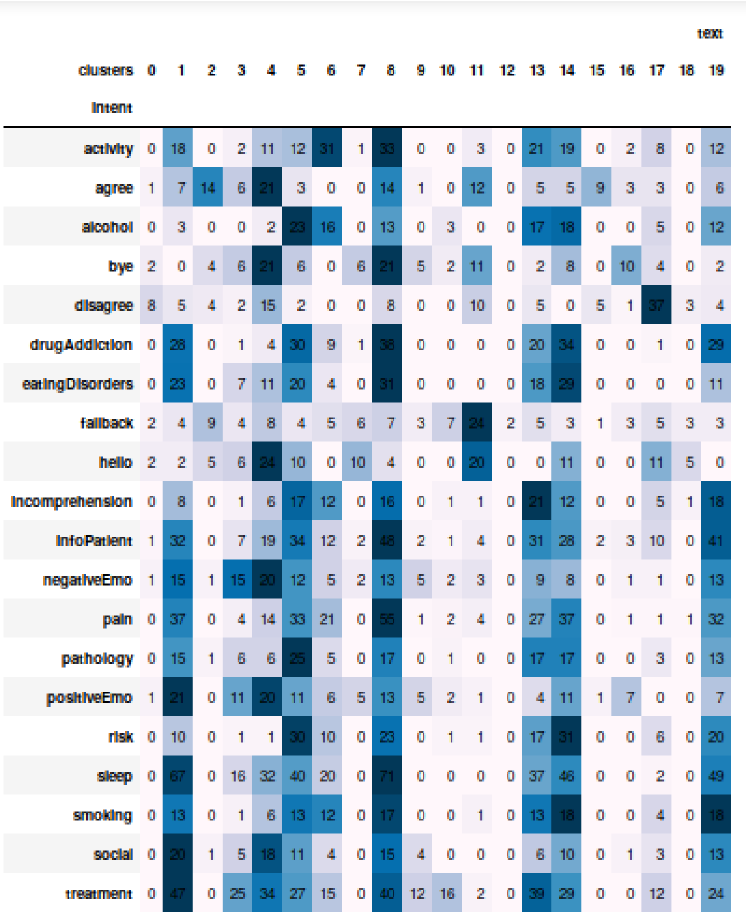


Figure 5: LDA scores



	text																			
clusters	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
intent																				
activity	0	18	0	2	11	12	31	1	33	0	0	3	0	21	19	0	2	8	0	12
agree	1	7	14	6	21	3	0	0	14	1	0	12	0	5	5	9	3	3	0	6
alcohol	0	3	0	0	2	23	16	0	13	0	3	0	0	17	18	0	0	5	0	12
bye	2	0	4	6	21	6	0	6	21	5	2	11	0	2	8	0	10	4	0	2
disagree	8	5	4	2	15	2	0	0	8	0	0	10	0	5	0	5	1	37	3	4
drugAddiction	0	28	0	1	4	30	9	1	38	0	0	0	0	20	34	0	0	1	0	29
eatingDisorders	0	23	0	7	11	20	4	0	31	0	0	0	0	18	29	0	0	0	0	11
fallback	2	4	9	4	8	4	5	6	7	3	7	24	2	5	3	1	3	5	3	3
hello	2	2	5	6	24	10	0	10	4	0	0	20	0	0	11	0	0	11	5	0
incomprehension	0	8	0	1	6	17	12	0	16	0	1	1	0	21	12	0	0	5	1	18
infoPatient	1	32	0	7	19	34	12	2	48	2	1	4	0	31	28	2	3	10	0	41
negativeEmo	1	15	1	15	20	12	5	2	13	5	2	3	0	9	8	0	1	1	0	13
pain	0	37	0	4	14	33	21	0	55	1	2	4	0	27	37	0	1	1	1	32
pathology	0	15	1	6	6	25	5	0	17	0	1	0	0	17	17	0	0	3	0	13
positiveEmo	1	21	0	11	20	11	6	5	13	5	2	1	0	4	11	1	7	0	0	7
risk	0	10	0	1	1	30	10	0	23	0	1	1	0	17	31	0	0	6	0	20
sleep	0	67	0	16	32	40	20	0	71	0	0	0	0	37	46	0	0	2	0	49
smoking	0	13	0	1	6	13	12	0	17	0	0	1	0	13	18	0	0	4	0	18
social	0	20	1	5	18	11	4	0	15	4	0	0	0	6	10	0	1	3	0	13
treatment	0	47	0	25	34	27	15	0	40	12	16	2	0	39	29	0	0	12	0	24

Figure 6: LDA confusion matrix

5.1.3 LDA model

num	purity	silhouette	homogeneity	completeness	AVG
10	0.253656	0.461589	0.157498	0.178928	0.262918
20	0.259203	0.378276	0.168191	0.182203	0.246968
30	0.251538	0.331069	0.157755	0.168448	0.227202
50	0.240545	0.287032	0.147430	0.166824	0.210458
100	0.266465	0.187596	0.168647	0.189221	0.202982
150	0.256077	0.167888	0.161589	0.185054	0.192652
200	0.263843	0.149212	0.175117	0.194913	0.195771
300	0.250025	0.151526	0.165607	0.184314	0.187868
400	0.272718	0.192777	0.177547	0.192270	0.208828
500	0.259506	0.203725	0.174925	0.192831	0.207747
550	0.229450	0.166200	0.149270	0.158548	0.175867
600	0.254261	0.191729	0.171744	0.186001	0.200934

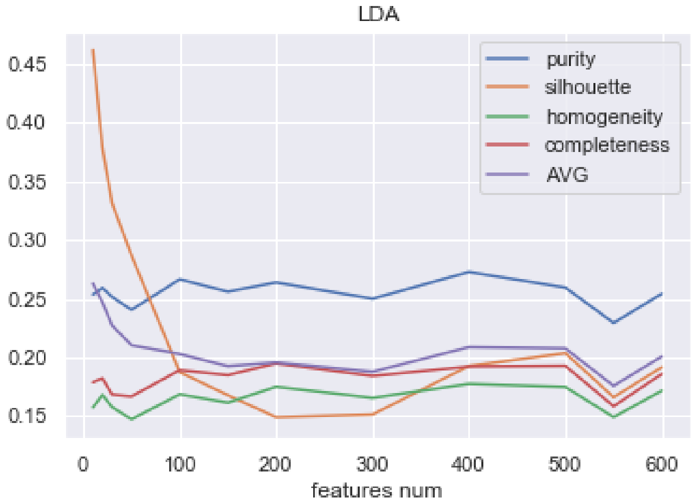


Figure 7: LDA scores

Best model - 500 topics.

5.1.4 GloVe pretrained

5.1.5 Google News W2V pretrained

We used word2vec google-news model with 300 dimensional feature space.

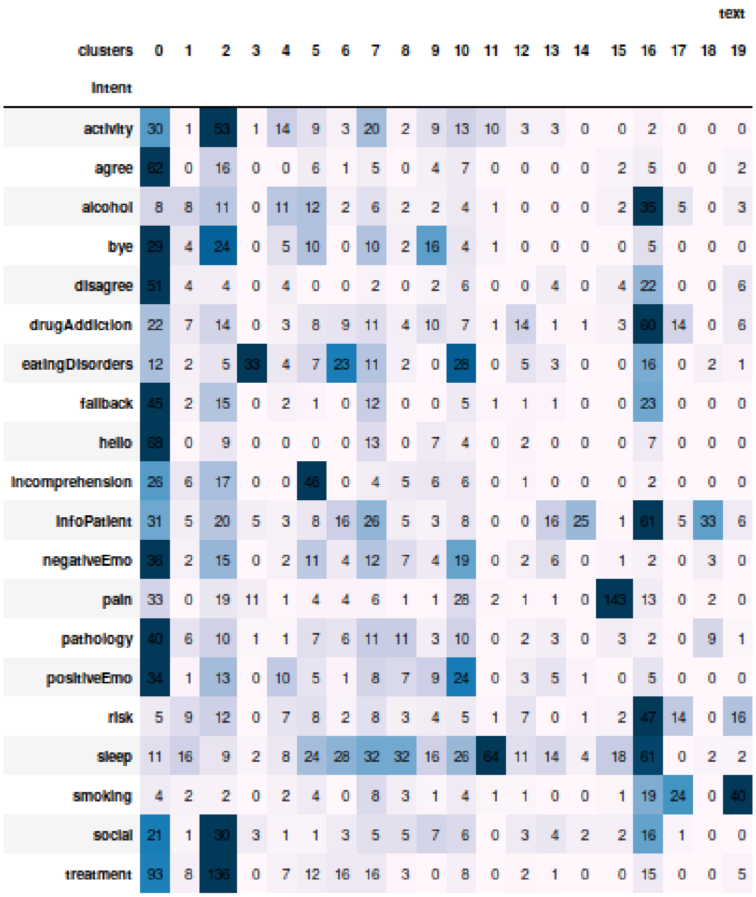
CV 0.418 + 0.057

5.1.6 CNN by word

From simple encoder, w2v, lda, w2v + lda

5.1.7 BiLSTM by word

From simple encoder, w2v, lda, w2v + lda



																				text
clusters	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
intent																				
activity	30	1	53	1	14	9	3	20	2	9	13	10	3	3	0	0	2	0	0	0
agree	62	0	16	0	0	6	1	5	0	4	7	0	0	0	0	2	5	0	0	2
alcohol	8	8	11	0	11	12	2	6	2	2	4	1	0	0	0	2	35	5	0	3
bye	29	4	24	0	5	10	0	10	2	16	4	1	0	0	0	0	5	0	0	0
disagree	51	4	4	0	4	0	0	2	0	2	6	0	0	4	0	4	22	0	0	6
drugAddiction	22	7	14	0	3	8	9	11	4	10	7	1	14	1	1	3	60	14	0	6
eatingDisorders	12	2	5	33	4	7	23	11	2	0	28	0	5	3	0	0	16	0	2	1
fallback	45	2	15	0	2	1	0	12	0	0	5	1	1	1	0	0	23	0	0	0
hello	68	0	9	0	0	0	0	13	0	7	4	0	2	0	0	0	7	0	0	0
incomprehension	26	6	17	0	0	46	0	4	5	6	6	0	1	0	0	0	2	0	0	0
infoPatient	31	5	20	5	3	8	16	26	5	3	8	0	0	16	25	1	61	5	33	6
negativeEmo	36	2	15	0	2	11	4	12	7	4	19	0	2	6	0	1	2	0	3	0
pain	33	0	19	11	1	4	4	6	1	1	28	2	1	1	0	143	13	0	2	0
pathology	40	6	10	1	1	7	6	11	11	3	10	0	2	3	0	3	2	0	9	1
positiveEmo	34	1	13	0	10	5	1	8	7	9	24	0	3	5	1	0	5	0	0	0
risk	5	9	12	0	7	8	2	8	3	4	5	1	7	0	1	2	47	14	0	16
sleep	11	16	9	2	8	24	28	32	32	16	26	64	11	14	4	18	61	0	2	2
smoking	4	2	2	0	2	4	0	8	3	1	4	1	1	0	0	1	19	24	0	40
social	21	1	30	3	1	1	3	5	5	7	6	0	3	4	2	2	16	1	0	0
treatment	93	8	136	0	7	12	16	16	3	0	8	0	2	1	0	0	15	0	0	5

Figure 8: LDA confusion matrix

5.1.8 Overall comparision

5.2 Classifier

5.2.1 Evoluation metrics

5.2.2 Results for each model

5.3 Semi-supervised learning

6 Conclusion

7 References

@InProceedingsN18-2058, author = ”Ferguson, James and Lockard, Colin and Weld, Daniel and Hajishirzi, Hannaneh”, title = ”Semi-Supervised

Event Extraction with Paraphrase Clusters”, booktitle = ”Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)”, year = ”2018”, publisher = ”Association for Computational Linguistics”, pages = ”359–364”, location = ”New Orleans, Louisiana”, url = ”http://aclweb.org/anthology/N18-2058”

@InProceedingssondhi-EtAl:2010:POSTERS, author = Sondhi, Parikshit and Gupta, Manish and Zhai, ChengXiang and Hockenmaier, Julia, title = Shallow Information Extraction from Medical Forum Data, booktitle = Coling 2010: Posters, month = August, year = 2010, address = Beijing, China, publisher = Coling 2010 Organizing Committee, pages = 1158–1166, url = http://www.aclweb.org/anthology/C10-2133

@articlearticle, author = Zhang, Thomas and H D Cho, Jason and Zhai, Chengxiang, year = 2015, month = 03, pages = , title = Understanding User Intents in Online Health Forums, volume = 19, journal = IEEE journal of biomedical and health informatics, doi = 10.1109/JBHI.2015.2416252

APA Zhang, T., Cho, J. H. D., & Zhai, C. (2015). Understanding User Intents in Online Health Forums. IEEE Journal of Biomedical and Health Informatics, 19(4), 1392-1398. [7066225]. https://doi.org/10.1109/JBHI.2015.2416252

Harvard Zhang, T, Cho, JHD & Zhai, C 2015, ’Understanding User Intents in Online Health Forums’ IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, 7066225, pp. 1392-1398. https://doi.org/10.1109/JBHI.2015.2416252