Automatic Information Extraction and Inferencing System from Online News Sources for Substance Abuse Cases

Judith George Joseph[a], Jestin Joy[b], Sreeraj M[c], Sanjay Govind[d], Shijas Muhammed T P[e] and Tibi Sunni[f]

[a]Department of Computer Science And Engineering, Federal Institute of Science And Technology(FISAT), Kerala, India
[b]Assistant Professor, Department of Computer Science And Engineering, Federal Institute of Science And Technology(FISAT), Kerala, India
[c]Assistant Professor, Department of Computer Science, Sree Ayyappa College, Alappuzha, Kerala
[d]Department of Computer Science And Engineering, Federal Institute of Science And Technology(FISAT), Kerala, India
[e]Department of Computer Science And Engineering, Federal Institute of Science And Technology(FISAT), Kerala, India
[f]Department of Computer Science And Engineering, Federal Institute of Science And Technology(FISAT), Kerala, India

Abstract

The rising number of substance abuse cases is a serious situation that demands significant attention. Gaining insights from the reported substance abuse cases will greatly help law enforcement authorities and policy makers. The unstructured nature of the publicly available data is a challenge. Computational techniques can be made use in efficiently extracting and summarising these unstructured data. The proposed system extracts the news reported on substance abuse related crimes from Malayalam online news papers. The extracted data is then processed using Natural Language Processing (NLP) techniques to generate a set of information that can be helpful in generating valuable inferences. Results show that the proposed system provide good accuracy for the data extraction task.

Keywords

Information extraction, NER, Machine Learning, Data Mining

1. Introduction

The United Nations Office on Drugs and Crime (UNODC) reports[1] that approximately 5 per cent of the world's population used an illicit drug in 2010 and 27 million people can be classified as problem drug users. Alcohol and illicit drug use cause around 39 deaths per million population. In addition to causing death, substance abuse is also responsible for significant morbidity and the treatment of drug addiction creates a tremendous burden on society. Significant rise in the reported drug abuse cases is a serious public health threat. Handling this problem needs the intervention of government, law enforcement and public health sector. World Health Organization (WHO) study[2] estimates that the four major cause of illicit drug use death are AIDS, suicide, overdose and trauma. Based on this, the median number of deaths are estimated to be 1,94,058 as per 2000 estimates. Illicit drug use also causes premature deaths in young adults and adversely affects their overall health.

Substance use is a problem in India too. Ministry of Social Justice and Empowerment, Government of India report[3] "Magnitude of Substance Use in India - 2019" shows the dismal picture in India. After Alcohol, Cannabis and Opioids is the most commonly used substances in India and about2.8% of the population use it. More than 30 lakh of the people with opioid use disorders are from Indian states of Uttar Pradesh, Punjab, Haryana, Delhi, Maharashtra, Rajasthan, Andhra Pradesh and Gujarat. Enforcement activities report[4] by Excise department, Government of Kerala reports that during 2019, 7099 cases are registered based on Narcotic Drugs and Psychotropic Substances Act.

Though governments publish[3, 4] data regarding substance abuse cases, it is not easy to get region wise detailed information. For example detailed information regarding size, type and location of registered cases are not easy to find. But these information are available in public domain through news reports. Problem with these news reports are that, they are not in a structured format. Various techniques[5, 6, 7] are explored for extracting structured information from unstructured textual data. Information extraction is

ISIC'21: International Semantic Intelligence Conference, February 25–27, 2021, New Delhi, India
judithgeorgejoseph123@gmail.com (J.G. Joseph); jestinjoy@fisat.ac.in (J. Joy); sreeraj.sac@gmail.com (S. M)
http://www.sreeayyappacollege.ac.in/uploads/downloads/sreeraj.pdf (S. M)
0000-0003-0892-7874 (J. Joy); 0000-0003-4974-437X (S. M)

the process of extracting information from unstructured data. It is extensively used in medical document mining, mining business and law documents. Internet being a rich source of unstructured textual data, web mining is also an active research area. The proposed system extracts structured information from news reports. News reports regarding substance abuse cases reported in online edition of popular Malayalam news papers are used for this purpose. These are then processed using Natural Language Processing (NLP) techniques like Named Entity Recognition (NER) for extracting structured information. This information helps in getting information like places where more cases are reported, most commonly used drug, amount of each drug as reported in news etc.

2. Related Works

Study on information extraction techniques from unstructured data is explored in literature[5, 6, 7, 8, 9]. This involves extracting data from medical text, business and law documents. Most of the research revolves around using English as the language. We haven't came across much research[10, 11] on information extraction from Malayalam unstructured text. This is mainly due to the unavailability of publicly available datasets and computational techniques for processing text. Works related to extracting drug related information unstructured text is discussed below.

Extracting Substance Abuse Information from Clinical Notes[8] was studied by Lybarger, Yetsigen et.al They proposed a neural network architecture for automatic extraction of substance abuse information from clinical notes. A discrete model was also experimented for extracting information. These clinical notes were stored with information about patients' substance abuse history. The model was trained to find the presence of substances events like alcohol, drug, or tobacco. A Maximum Entropy (MaxEnt) model was used for classifying the status. Other entities like *amount, frequency, exposure history,..* were extracted using Conditional random fields (CRF) model. Neural Multi-task Model predicted all entities for all substances.

Khmael Rakm Rahem and Nazlia Omar proposed a rule-based approach [9] for extracting drug related crime information from online newspaper articles. The task involved extracting information like drug name, nationality, location and assess the quantity and price of drug. A set of grammatical and heuristic rules were used for this purpose. Data from Malaysian National News Agency (BERNAMA) is used in the system. System achieved a precision of 0.96 on drug names, 0.83 on crime location and 0.87 on drug quantity. But information regarding the dataset size and testing information is missing in the paper.

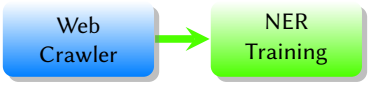

Figure 1: NER Training

Rexy Arulanandam, Bastin Tony Roy Savarimuthu and Maryam A. Purvis proposed a system[12] for extracting crime information from newspaper articles. Named Entity Recognition (NER) coupled with Conditional Random Field (CRF) is used to find crime location in a sentence. 70 articles from Otago Daily Times is used for evaluating the system. LBJ NER Tagger is found to be the best tagger with a precision of 0.98. Accuracy varies from 84% to 90% for for New Zealand articles for the task of identifying locations in sentences and classifying it into crime location sentences.

Eiji Aramakia, Yasuhide Miurab, Masatsugu Tonoike et al[13] proposed a system for extracting adverse drug events and effects from clinical records. Results on a study on 3,012 discharge summaries show that 7.7% of records include adverse event information, and 59% of them can be extracted automatically.

Authors haven't came across any similar systems for extracting information from Malayalam news articles.

3. Design and Implementation

Proposed system consists of two phases. In the first phase, relevant data is crawled from web and fed to Named Entity Recognition Module (NER) for creating a model for recognizing named entities. This phase is given in Figure 1.

This phase is not an easy task since we need to NER on Malayalam language text. Malayalam[14] is a language spoken in the Indian state of Kerala. It is one of 22 scheduled languages of India and is spoken by 37,919,870 people. Malayalam follows a word order of SOV (*subject-object-verb*) generally. Malayalam is a heavily agglutinated and inflected language making it difficult for NER task. Different techniques are explored for Malayalam NER[15, 16, 17, 18, 19]. Most of these are based statistical techniques. This study also used a statistical technique for NER.

Statistical model provided by Spacy[1] is used in this study. Tagged data is fed to the NER system for train-

[1]https://spacy.io/

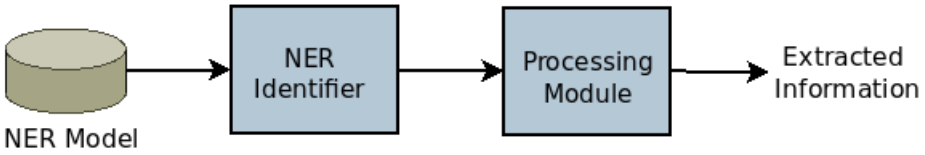

Figure 2: Information extraction

ing. Transition based approach[20] is used for NER. This uses word embedding strategy using subword features and bloom embeddings. CNN filter sizes are chosen with beam search. 1D convolutional filters are applied over the input text to predict how the upcoming words may change the current entity tags.

In the second phase, the trained model is made use in extracting information. A rule base is also used for this purpose. This is given in Figure 2. NER model helps to identify relevant entities for the information extraction task. *Name, age, place, drug name, amount* and *size* is considered in the proposed system. Tagged sentences are then fed to processing module, which processes information based on handcrafted rules. A snapshot of the rules used in the proposed system are given below.

1. *Name of the person, and drug appears in the initial part of the news item.*
2. *If money occurs just before the drugs name, then it is assigned as that of the corresponding drug's*
3. *First occurrence of location is assigned as that of location*
4. *Person age is close to the name of the person*
5. *Amount of drug carried by the offender is close to the drug name*

3.1. Dataset

Though there exists trained models for languages like English, publicly available tagged dataset for Malayalam language is non existent for this task. Data is extracted from online edition of Malayalam news sites of Malayala Manorama, Mathrubhumi, Mangalam, News18 Malayalam, Deshabhimani and Media One. Tagging for NER was done using web frontend based on doccano[21]. It is an open source text annotation tool. It can be used to tag data for various tasks like named entity recognition, text summarization and sentiment analysis. Data collected for training were from the period January 2017 to December 2019.

കാസർകോട്, എംഡിഎംഎ, നള്ളിപ്പാടിയിലെ, കാസർകോട്, മെൽവിൻ ജോസിന്റെ, ചെയ്ത്.,ചൊവ്വാഴ്ച, വൈകിട്ട് നാലുമണിയോടെ, നള്ളിപ്പാടിയിൽ, റബിയത്തിനെ, 250 ഗ്രാം, മയക്കുമരുന്ന്, ബംഗളൂരുവിൽ, മയക്കുമരുന്നെത്തിക്കുന്ന, റാബിത്തെന്ന്

Location കാസർകോട്
drug എംഡിഎംഎ
Location നള്ളിപ്പാടിയിലെ
Location കാസർകോട്
Person മെൽവിൻ ജോസിന്റെ
Date ചെയ്ത്.,ചൊവ്വാഴ്ച
Time വൈകിട്ട് നാലുമണിയോടെ
Location നള്ളിപ്പാടിയിൽ
Person റബിയത്തിനെ
Quantity 250 ഗ്രാം
drug മയക്കുമരുന്ന്
Location ബംഗളൂരുവിൽ
drug മയക്കുമരുന്നെത്തിക്കുന്ന
Person റാബിത്തെന്ന്

കൽപ്പറ്റ, മുത്തങ്ങ, ഹാൻസ്, 15000 പായ്ക്കറ്റ്, ഹാൻസാണ്, വെണ്ണക്കാട്, സലീം

Location കൽപ്പറ്റ
Location മുത്തങ്ങ
drug ഹാൻസ്
Quantity 15000 പായ്ക്കറ്റ്
drug ഹാൻസാണ്
Location വെണ്ണക്കാട്
Person സലീം

Figure 3: NER output for processed sample news items

3.2. Implementation

Processing of the data is done using Python programming language. Spacy[2] NER module is used for named entity recognition, which forms the important component of the system. The availability of pretrained statistical models and support for large number of languages makes Spacy a good choice for text processing.

4. Results and discussion

The proposed system involves passing the news item to NER module and processing it using the rule based system. Figure 3 shows the result of NER module for sample news items.

This is then fed to the processing module for inference. Output from the inference module is given in Figure 4.

[2]https://spacy.io/

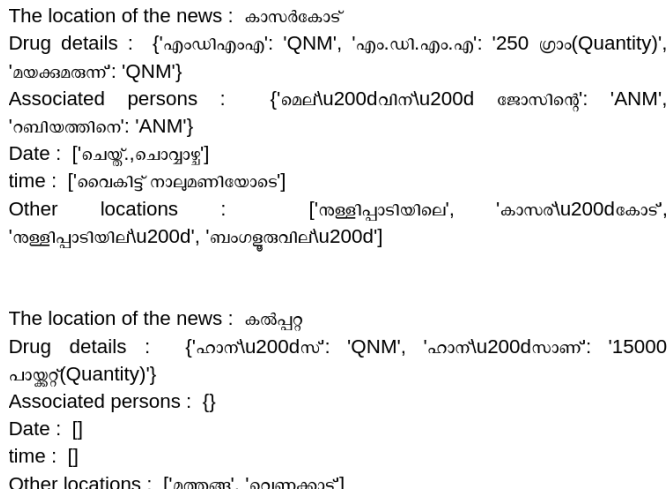

Figure 4: Output of inference module

Entity	Correctly Identified(50)	Accuracy
Location	42	0.84
Drugs	40	0.80
Quantity	30	0.60
Money	38	0.76
Person	30	0.60
Age	32	0.64
Date	30	0.60
Time	36	0.72

Table 1
Accuracy of each entity tested 50 news articles

For evaluating the system, 50 substance abuse related news articles are collected. These news articles are from the period January 2020 to March 2020. The collected news articles are manually verified to be of substance abuse cases. These news articles are then fed to system and accuracy of the entity identified is recorded. Accuracy is found by matching the entities manually with the predicted entities. For example results indicate that of the 50 news articles considered, on 42 of them location entity is predicted correctly.

Table 1 lists the accuracy identified by the system for the given 50 news articles.

Results indicate that system could identify the entities *location, drugs* with reasonably good accuracy. Although system could identify most entities correctly, these are marked as those relevant by the rule based system. The reduction in accuracy for other entities is due to the failure in the part of rule base to correctly match the entity. Rules are framed manually after going through news stories. We cant be sure of each and every news story following the same writing style. This is a major drawback of the system. For example the accuracy of the entities *quantity, person, date* are the lowest. Most news stories lack *quantity* and *date* information in a standard format. *Person* information is also difficult to identify since news stories sometimes lack them and sometimes more person names like that of law enforcement authorities are included making it difficult for the system to correctly identify it.

5. Conclusion

An automated system for generating valuable information out of online news articles can reduce the colossal amount of effort that must be put in to do the same by other means. The data provided by the system can aid in statistical research and study, generating key inferences for investigations, for background studies in formulating action plans etc. Since the system processes news reports on crimes related to substance abuse, the information provided is very significant and relevant as the issue is an ongoing serious social threat.

However in a broad sense the services provided by the current version of the system is limited. Which also opens an opportunity for future enhancement. Now the system is providing only key aspects mentioned in news. It can be modified into a full fledged inference which increase it's clarity. The proposed system can be enhanced in a way that it responds to user queries.

Acknowledgments

Authors would like to thank the help extended by Adam Shamsudeen for providing the required dataset and tagging frontend.

References

[1] UNODC, Atlas on substance use (2010), 2011. URL: https://www.who.int/publications/i/item/9789241500616.

[2] M. W.-S. Louisa Degenhardt, Wayne Hall, M. Lynskey, Illicit drug use, 2020. URL: https://www.who.int/publications/cra/chapters/volume1/1109-1176.pdf.

[3] N. D. D. T. C. (NDDTC), Magnitude of substance use in india - 2019, 2020.

[4] G. o. K. Excise department, Month wise details of enforcement activities during 2019, 2020. URL: https://excise.kerala.gov.in/enforcement-activities-2/.

[5] M. Alawad, S. Gao, J. X. Qiu, H. J. Yoon, J. Blair Christian, L. Penberthy, B. Mumphrey, X.-C. Wu, L. Coyle, G. Tourassi, Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks, Journal of the American Medical Informatics Association 27 (2020) 89–98.

[6] S. Jiang, S. Baumgartner, A. Ittycheriah, C. Yu, Factoring fact-checks: Structured information extraction from fact-checking articles, in: Proceedings of The Web Conference 2020, 2020, pp. 1592–1603.

[7] N. Milosevic, C. Gregson, R. Hernandez, G. Nenadic, A framework for information extraction from tables in biomedical literature, International Journal on Document Analysis and Recognition (IJDAR) 22 (2019) 55–78.

[8] K. Lybarger, M. Yetisgen, M. Ostendorf, Using neural multi-task learning to extract substance abuse information from clinical notes, in: AMIA Annual Symposium Proceedings, volume 2018, American Medical Informatics Association, 2018, p. 1395.

[9] K. R. Rahem, N. Omar, Drug-related crime information extraction and analysis, in: Proceedings of the 6th International Conference on Information Technology and Multimedia, IEEE, 2014, pp. 250–254.

[10] N. Mohandas, J. P. Nair, V. Govindaru, Domain specific sentence level mood extraction from malayalam text, in: 2012 International Conference on Advances in Computing and Communications, IEEE, 2012, pp. 78–81.

[11] D. S. Nair, J. P. Jayan, E. Sherly, et al., Sentimasentiment extraction for malayalam, in: 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2014, pp. 1719–1723.

[12] R. Arulanandam, B. T. R. Savarimuthu, M. A. Purvis, Extracting crime information from online newspaper articles, in: Proceedings of the second australasian web conference-volume 155, 2014, pp. 31–38.

[13] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, K. Ohe, Extraction of adverse drug effects from clinical records., MedInfo 160 (2010) 739–743.

[14] G. F. Simons, C. D. Fennig, Ethnologue: languages of Asia, sil International Dallas, 2017.

[15] P. Sreeja, A. S. Pillai, Towards an efficient malayalam named entity recognizer analysis on the challenges, Procedia Computer Science 171 (2020) 2541–2546.

[16] C. Malarkodi, S. L. Devi, A deeper study on features for named entity recognition, in: Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation, 2020, pp. 66–72.

[17] J. P. Jayan, R. Rajeev, E. Sherly, A hybrid statistical approach for named entity recognition for malayalam language, in: Proceedings of the 11th Workshop on Asian Language Resources, 2013, pp. 58–63.

[18] A. Ajees, S. M. Idicula, A named entity recognition system for malayalam using neural networks, Procedia computer science 143 (2018) 962–969.

[19] S. Thottingal, Finite state transducer based morphology analysis for malayalam language, in: Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, 2019, pp. 1–5.

[20] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360 (2016).

[21] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, X. Liang, doccano: Text annotation tool for human, 2018. URL: https://github.com/doccano/doccano, software available from https://github.com/doccano/doccano.