Information Extraction for Additive Manufacturing Using News Data

Neha Sehgal[1,2](✉) and Andrew Crampton[1]

[1] University of Huddersfield, Queensgate, Huddersfield, UK
[2] Valuechain, 3MBIC, Huddersfield, UK
nsehgal@valuechain.com

Abstract. Recognizing named entities like Person, Organization, Locations and Date are very useful for web mining. Named Entity Recognition (NER) is an emerging research area which aims to address problems such as Machine Translation, Question Answering Systems and Semantic Web Search. The study focuses on proposing a methodology based on the integration of an NER system and Text Analytics to provide information necessary for business in Additive Manufacturing. The study proposes a foundation of utilizing the Stanford NER system for tagging news data related to the keywords "Additive Manufacturing". The objective is to first derive the organization names from news data. This information is useful to define the digital footprints of an organization in the Additive Manufacturing sector. The existence of an organization derived using the NER approach is validated by matching their names with companies listed on the Companies House portal. The organization names will be matched using a Fuzzy-based text matching algorithm. Further information on company profile, officers and key financial data is extracted to provide information about companies interested and working within the Additive Manufacturing sector. This data gives an insight into which companies have digital footprints in the Additive Manufacturing sector within the UK.

Keywords: Named Entity Recognition · News data · Additive Manufacturing · Text matching · Open data

1 Introduction

Additive Manufacturing has the potential to revolutionize the global parts manufacturing and logistics landscape in the UK. It enables distributed manufacturing and the production of parts-on-demand while offering the potential to reduce cost and energy consumption; and thereby carbon footprint. This paper explores a paradigm data science approach to gather information on companies associated with Additive Manufacturing to fully exploit Additive Manufacturing growth and potential. There is no special SIC Code associated with companies

Supported by Knowledge Transfer Partnership, Innovate UK.

H. A. Proper and J. Stirna (Eds.): CAiSE 2019 Workshops, LNBIP 349, pp. 132–138, 2019.
https://doi.org/10.1007/978-3-030-20948-3_12

related to Additive Manufacturing sector, therefore it is vital to derive the list of companies, engaged with Additive Manufacturing, either digitally or registered with the UK Companies House portal. To find such companies, this paper proposes to utilize news data to find the organization being discussed in recent news articles.

In this study, the news data is collected for the keyword "Additive Manufacturing" from different open API sources. The task of extracting organizations and names of people from textual data is known as Named Entity Recognition (NER). Named entity recognition tasks generally require costly, hand-labelled training data and most existing corpora are small in size. Therefore, it is better to use the Stanford NER system to extract the names of persons and organizations listed in news articles for the keyword "Additive Manufacturing". Further, text matching is performed to match organizations derived by the NER system with companies registered at Companies House Portal.

2 Background

2.1 Named Entity Recognition

This section covers the landscape of various state-of-the-art approaches for Named Entity Recognition and Classification tasks. Historically NER systems were based on hand-made rules, but, due to the growth of big data and the popularity of machine learning in recent times, researchers have developed powerful, reliable and robust NER systems. The work related to NER has been studied extensively and considers major factors such as: Language, Textual genre and Entity type [1]. The ability to learn from previously known entity data is an essential part of any NER problem. The words, with their associated tags, compose the feature set for supervised learning.

Recent studies have utilized supervised machine learning to find insights from training data and induce rule-based systems. Major supervised learning approaches include Hidden Markov Models (HMMs), Ensemble Models, Support Vector Machines and Artificial Neural Networks. Su *et al.* [2] propose a Hidden Markov Model-based chunk tagger to recognize and classify names, times and numerical data. [3] employs a maximum entropy concept to use global information directly for the NER task. [4] experimented with a combination of four diverse classifiers for the NER task, including linear classifiers, maximum entropy, transformation-based learning and HMMs. [5] proposed a probabilistic approach, combined with Latent Dirichlet Allocation, to employ supervised learning using partially labeled seed entities. A study by [6] used Support Vector Machines for feature selection and for the Named Entity recognizer task.

Few researchers have employed unsupervised learning, i.e. clustering to gather named entities from clusters created based on their similarity of context [7]. Recent examples of NER applications include monitoring Twitter streams to understand user's opinions and sentiments. Li [7] presented a novel, unsupervised NER system for Twitter streams using dynamic programming followed by a random walk model. Ritter *et al.* [8] developed a novel T-NER system which outperformed the Stanford NER System in terms of the F1-Score by overcoming redundancy inherent in tweets using LabeledLDA.

With the rise in availability of social media data, it is difficult to analyze data, in the form of news, due to challenges including variations in spelling, linguistics, commenting, emoticons, images and the use of mixed languages. As it is difficult to annotate text data from new databases, due to time and cost constraints, it is beneficial to use the Stanford NER system (an already annotated corpus) to derive the organization being discussed and highlighted in discussions related to Additive Manufacturing. Many NER-related studies [9,10] have shown that Stanford NER generally performs better than other corpora. Therefore, this study plans to use Stanford NER for tagging and extracting organizations from corpora of news data.

3 Dataset

In order to build a picture of the Additive Manufacturing sector in the UK, data has been gathered and aggregated from multiple news API for the keywords "Additive Manufacturing". In total, 10,000 news articles are extracted which include headlines, description, paragraphs, author name and any associated tags.

The information on matched company data, after applying the NER system and text matching, has been gathered from Companies House. This information has been captured for all companies which are listed as active as of the 1st January 2019, who have a Manufacturing SIC Code. The data on company profile, officers and financial information is extracted using the Companies House API by using Python. The financial data provides details regarding the company's balance sheet, together with a few more important parameters.

Using a custom algorithm, potential company websites have been identified and scraped for: self-reported descriptions of the company; information about the sectors that the company reports they operate in; the accreditations they report that they hold; information about their capabilities; and additional links to social media accounts. This data gives an insight into which companies have digital footprints within the Additive Manufacturing sector in the UK.

3.1 Tools

Currently, there are vast amounts of tools available for data analysis and machine learning. Python has been ranked number one for the last two years on LinkedIn for being the most powerful and popular tool for data science. Python is chosen in this study as the programming language of choice due to its seemingly vast adoption in the data analysis and machine learning community.

Three major Python libraries i.e. NumPy, Panda and Scikit-learn will be used for data analysis. Panda will be used for checking the types of attributes and for relevant modifications as necessary. NumPy is useful for converting dataframes to array formats which is essential for modeling purposes. Scikit-learn is useful for clustering tasks and for splitting data into training, validation and testing groups.

4 Proposed Methodology

The news data is extracted from different news APIs. After the data has been extracted, the text data is cleaned by removing punctuation, transforming to lower case and by removing stop words. The Standford NER system is then implemented to extract names of organizations as follows:

4.1 Entity and Relation Extraction

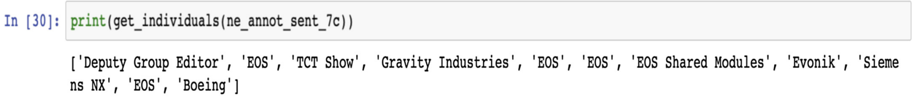
The StanfordNER Tagger, in the NLTK library, is used for extracting the Named Entity Recognition which is a sequence of words (news data) consisting of: Name **(PERSON, ORGANIZATION and LOCATION)**; Numerical **(MONEY and PERCENTAGE)**; and Temporal **(DATE and TIME)**.

- For extracting the NER, we used the NER Model trained on an English corpus i.e., **“english.muc.7class.distsim.crf.ser”**
- For extracting the NER, we also used an NER Tagger engine i.e., **“stanford-ner.jar”** (it is also know as CRF classifier).

For applying the NER Tagger we have two techniques available:

1. Pass the refined text, obtained previously, as a parameter to the Stanford NER Tagger to get the NER tags.
2. Convert the refined text into sentences and send these as a parameter to the Stanford NER Tagger to get the NER tags.

A snapshot of the list of organizations from a news article is shown below in Fig. 1.



```
In [30]: print(get_individuals(ne_annot_sent_7c))

         ['Deputy Group Editor', 'EOS', 'TCT Show', 'Gravity Industries', 'EOS', 'EOS', 'EOS Shared Modules', 'Evonik', 'Sieme
         ns NX', 'EOS', 'Boeing']
```

Fig. 1. Grouped NER tags on the basis of (ORGANIZATION)

The extracted names of organizations from 10,000 news articles are validated with companies registered on the Companies House Portal using a fuzzy-based text matching algorithm. The success rate for validation is 70% i.e., approx. 70% of company names extracted from online news articles had matched with Companies House data. Lastly, the data for company profiles, officers and financial information is extracted from the Companies House API. The data provides a holistic overview about UK companies working or interested in Additive Manufacturing. This data is important for start-up companies, investors and SMEs in order to understand the progress encompassing the Additive Manufacturing sector in UK.

The Companies House data on companies extracted through this methodology provides information on digital footprints and financial performance over a number of years.

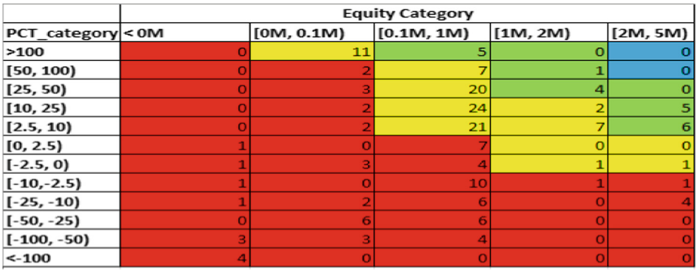
5 Results

The results will be discussed in detail during the presentation. Due to data privacy, the whole analysis is not presented here in the paper. To provide an overview, a sample of 515 news articles on the topic of Additive Manufacturing are selected randomly from the corpus. The proposed methodology, based on NER, is applied to derive the list of organizations discussed in 515 news articles. There were a total of 3175 organization names extracted from 515 news articles. A sample of 750 Small and Medium Enterprises (SME) from 3175 organization names was selected for further information and data analysis. The company names are matched with a list of companies listed in the Companies House database using a fuzzy-based text matching algorithm. For this sample data, approximately 43.5% of company's names, found using the NER approach, matched with Companies House data.

The data on company profile, officers and financial (equity, assets and liabilities) are gathered from the Companies House API for 327 companies. The financial data was extracted for 192 SME companies. The remaining 135 companies are either start-up, or dormant companies or they have filed financial details in paper format; in which case their financial details are not available.

The company profile data shows that approximately 34% of additive manufacturing companies falls within the Greater Manchester Local Enterprise Partnership (LEP). The company age profile shows that 11% of additive manufacturing SMEs have recently established with an age less than 2 years. The age distribution of directors associated with additive manufacturing SME Companies shows that the majority of directors falls within the category of 45–55 years of age.

The companies are further classified based on their balance sheet banding and percentage change in shareholder funds over the last two years. The companies are categorized in four groups: Champions (Blue), Contenders (Green), Prospects (Yellow), and Strugglers (Red). The data shows that there are a high number of Prospects SME in the additive manufacturing sector. The financial matrix, as shown in Fig. 2, is helpful for investors and other key players of manufacturing sectors, looking for opportunity to expand businesses or connect with members of supply chain.



	Equity Category				
PCT_category	< 0M	[0M, 0.1M)	[0.1M, 1M)	[1M, 2M)	[2M, 5M)
>100	0	11	5	0	0
[50, 100)	0	2	7	1	0
[25, 50)	0	3	20	4	0
[10, 25)	0	2	24	2	5
[2.5, 10)	0	2	21	7	6
[0, 2.5)	1	0	7	0	0
[-2.5, 0)	1	3	4	1	1
[-10,-2.5)	1	0	10	1	1
[-25, -10)	1	2	6	0	4
[-50, -25)	0	6	6	0	0
[-100, -50)	3	3	4	0	0
<-100	4	0	0	0	0

Fig. 2. Financial analysis for SME in Additive Manufacturing (Color figure online)

6 Conclusion

In the era of machine learning and big data, information extraction plays an important role for parsing and classifying billions of news articles. If the category of any company specification is not listed as one of the key SIC Codes, an alternative approach can be considered by integrating NLP and text matching algorithms. Information about organizations working or interested in the Additive Manufacturing sector can be gathered by using news articles listing the keywords "Additive Manufacturing". NER systems can help in extracting organization names from news data, which has been validated with Companies House data and thereby provides more detailed knowledge about companies by gathering information on its profile, officers and financial situation.

References

1. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investig. **30**(1), 3–26 (2007)
2. Zhou, G.D., Su, J.: Named entity recognition using an HMM-based chunk tagger. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002), pp. 473–480. Association for Computational Linguistics, Stroudsburg (2002). https://doi.org/10.3115/1073083.1073163
3. Chieu, H.L., Ng, H.T.: Named entity recognition: a maximum entropy approach using global information. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), vol. 1, pp. 1–7. Association for Computational Linguistics, Stroudsburg (2002). https://doi.org/10.3115/1072228.1072253
4. Florian, R., Ittycheriah, A., Jing, H., Zhang, T.: Named entity recognition through classifier combination. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 (CONLL 2003), vol. 4, pp. 168–171. Association for Computational Linguistics, Stroudsburg (2003). https://doi.org/10.3115/1119176.1119201
5. Guo, J., Xu, G., Cheng, X., Li, H.: Named entity recognition in query. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009), pp. 267–274. ACM, New York (2009). https://doi.org/10.1145/1571941.1571989
6. Isozaki, H., Kazawa, H.: Efficient support vector classifiers for named entity recognition. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), vol. 1, pp. 1–7. Association for Computational Linguistics, Stroudsburg (2002). https://doi.org/10.3115/1072228.1072282
7. Li, C., et al.: TwiNER: named entity recognition in targeted Twitter stream. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012), pp. 721–730. ACM, New York (2012). https://doi.org/10.1145/2348283.2348380
8. Ritter, A., Clark, S., Mausam, Etzioni, O.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), pp. 1524–1534. Association for Computational Linguistics, Stroudsburg (2011)

9. Kazama, J.I., Torisawa, K.: Exploiting Wikipedia as external knowledge for named entity recognition. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (2007)
10. Nothman, J., Curran, J.R., Murphy, T.: Transforming Wikipedia into named entity training data. In: Proceedings of the Australasian Language Technology Association Workshop 2008, pp. 124–132 (2008)