Social Media Mining for Ideation by using Classification Methods

Sercan Ozcan sercan.ozcan@port.ac.uk
Portsmouth Business School, University of Portsmouth, Portsmouth
Sushant Chatufale sushantchatufale@gmail.com
University of Portsmouth
C. Okan Sakar okan.sakar@eng.bau.edu.tr
Department of Computer Engineering, Bahcesehir University

ABSTRACT

Ideation is the most crucial and the first step of almost any innovation process [1]. It starts with either identifying a problem or creating a need for any product/service development process. Ideation process can occur within or outside of the organisations' endeavours. Nowadays, many firms have procedures and software to motivate and collect ideas internally. This is a very common approach for especially large companies as they do have an access to enormous human capital and resources. Considering external resources, open innovation, co-creation or crowdsourcing are popular concepts and approaches where companies interact with consumers, inventors and other organisations to enhance their innovation capability.

One of the most popular process in all open external approaches is to gather information from consumers for different stages of the innovation process. Many companies have structured systems, methods and targeted competitions to obtain valuable inputs from consumers about their existing products or potential innovations. One innovative approach and a cost-effective alternative to collect ideas from external resources could be based on a social media data such as Twitter, Facebook and Instagram. Social media is a great source to mine data especially if the required information is about consumers and products.

Twitter is one of the top social media platform where users can exchange short messages (tweets) up to 140 characters and can follow other users to be notified of their activity [2]. These tweets are shared with specific hashtags and half a billion tweets are generated every day from millions of Twitter users. Using social media based big data, it is possible to obtain information about consumers, trends, companies and technologies using text mining techniques. However, the data quality is one significant concern when it comes to the social media based analysis, especially in cases where semantic, text mining or natural language processing related approaches are used to retrieve information. The quality of the retrieved data from the social media can be enhanced if it is retrieved from specific resources with optimised techniques.

This paper aims to mine Twitter data to explore the trends and retrieve ideas for different purposes such as product development, technology and sustainability oriented considerations. The main approach of this study is to classify the tweets to be an idea or not. These retrieved ideas provide insights about expectations, problems or needs of consumers and organisations. The results also illustrates the reactions of consumers for technological developments. Tweets are retrieved based on specific hashtags using Twitter Search API. Various supervised and unsupervised classification algorithms are used to classify if a tweet is an idea or not. The classification algorithms are also compared for various validation metrics.

This study retrieves Twitter data from 2016 to 2018 containing different combinations of the hashtags #idea, #technology, #sustainability and #npd. Twitter data is mined to also illustrate the sector specific information such as the statistical results related to the distribution of ideas [3]. The sectors which have large number of ideas is considered for further sector based specific analysis. The sector based selection is used to increase the accuracy of the classification results.

The data is retrieved in the form of tweets from Twitter based on what consumers are saying related to ideas, technology, new product development (NPD) and sustainability.

After data retrieval, the data is cleaned to eliminate grammatical and spelling mistakes, and to eliminate unwanted URLs and characters using data cleaning and pre-processing methods such as stop-word removal, regular expressions, and lemmatisation. The cleaned tweets are categorised using classification methods such as SVM, Random Forest and Neural Networks [4]. For the semi-supervised method, we labelled some of our data using clustering and principal component analysis (PCA) techniques, while we also use a manually annotated dataset for classifying the tweets as "idea" or "no idea". A comparison of these classification algorithms is completed to identify the highest classification accuracies. The results are visualised using a word cloud method to show popular words in tweets containing idea hashtags.

The results demonstrate that our method based on text mining and classification methods can extract ideas from consumers and is a great method to show technological trends [5]. In addition, this study illustrates the conditions where semi-supervised or unsupervised classification methods work the best. The quality and the accuracy of the results are increased when the data is retrieved from combination of hashtags and the classification methods are optimised for these specific hashtags. Companies and entrepreneurs can use this method to identify information for their product development activities.

References:

[1] Ende, J., Frederiksen, L., & Prencipe, A. (2015). The front end of innovation: Organizing search for ideas. Journal of Product Innovation Management, 32(4), 482-487.

[2] Balachander Krishnamurthy, Phillipa Gill , Martin Arlitt, A few chirps about twitter, Proceedings of the first workshop on Online social networks, August 18-18, 2008, Seattle, WA, USA [doi>10.1145/1397735.1397741]

[3] Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In Proceedings of the First Workshop on Social Media Analytics (SOMA '10). ACM, New York, NY, USA, 80-88. DOI=http://dx.doi.org/10.1145/1964858.1964870

[4] J. Li, H. Xu, X. He, J. Deng and X. Sun, "Tweet modeling with LSTM recurrent neural networks for hashtag recommendation," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 1570-1577. doi: 10.1109/IJCNN.2016.7727385

[5] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal and A. Choudhary, "Twitter Trending Topic Classification," 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, 2011, pp. 251-258. doi: 10.1109/ICDMW.2011.171