SEUSL

Sri Lankan Journal of Technology

Research
Open Access

SLJoT
2021 Sp Issue 67-72
ISSN 2773-6970

Product Attribute Extraction from C2C Social Media Messages

Mohamed Refai Mohamed Rilfi

Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka

[*]Corresponding Author: rilfi@cse.mrt.ac.lk

Received: 05-05-2021 ∗ Accepted: 14-07-2021 ∗ Published Online: 15-09-2021

***Abstract*—**On social media, people could share information related to their desire to purchase, sell, or consume products or services, which serves as a marketplace for C2C e-Commerce. However, the message post by the social media users will not reach the potential buyer/seller out of your followers' circle. Furthermore, due to the difficulties of interpreting the semantics of social media posts, extracting product attribution from them is also difficult. To fix these issues, our research proposes a framework for extracting product attributes from microblogging messages about product selling and buying in this paper. First, we use a hybrid approach that includes Knowledge Base (KB), rule-based, Conditional Random Field (CRF), and Logistic Regression to extract the semantics of messages using named entity recognition. The dataset was created using raw social media messages, product descriptions from e-commerce sites, and KB because there was no product attribute annotated training dataset. When applied to a real-world dataset, the proposed approach achieves high accuracy, with classification and CRF models achieving 95 and 82 percent accuracy, respectively.

***Keywords*—**C2C, social media stream, knowledge base, information extraction, named entity recognition

I. Introduction

The consumer-to-consumer (C2C) business model obviates the need for middlemen between sellers and buyers(Dan, 2014). Social networking is increasingly being used by C2C e-commerce as a free and real-time medium for communication between potential buyers and sellers. There are so many messages on social networking sites, as well as sharing procedures slanted towards distorted social ties, that it is nearly impossible to locate a needle in a haystack. Although automation is the answer, these messages, like most social media posts, are short, informal, unstructured, and even cryptic, making them difficult to understand. Furthermore, there is a great deal of variety in terms of product and service details.

Although there is no single solution to the C2C product attribute extraction problem, there are many options that address various aspects of the problem. For example, in (Raju, Pingali, and Varma, 2009), the authors proposed using n-gram classification to extract product attributes using an unsupervised machine-learning technique. However, their research was focused on pre-processing and discovering string similarities between entities that belonged to the same product category but had slight differences due to human error. Furthermore, when compared to the Conditional Random Fields (CRF) algorithm, the accuracy of the defined attributes is poor (Bing et al., 2016) and describes an unsupervised method for extracting popular product attributes from e-commerce websites.

Real-time knowledge extraction using Twitter (Nozza et al., 2017) is by some means very relevant to our research but They using geographical location data as their primary information retrieval searching terms to collect the bulk of messages from Twitter. They are implementing centralized social media processing in this research. Central processing is subjected to failure. Other than for real-time data processing, there are no results that can be given. For data extraction, he applied the knowledge base (Derczynski et al., 2015) to obtain the correct keywords, but named entities were unnecessary. Ahmad, in a doctoral dissertation presented text messaging research, which demonstrated how difficult it is to extract information from a text message. They have placed their attention on emotion analytics. Emojis, in particular, are analyzing the text messages using the SVM (Burges, 1998) and KNN (Han, Karypis, and Kumar, 2001) algorithm. Clustering for personalization (Jenhani et al., 2018) concentrating on social media dissemination, the redistribution of computations across multiple networks. Apache Storm used stream processing with low latency to produce near-real-time output by looking at the stream performance analysis report, an increase in throughput was observed when using a distributed processing strategy.

The different forms of real-time text stream processing were implemented to identify different types of crises using unregulated social media (Tarasconi et al., 2017). When they first started developing their filters, they only filtered social media content. Combining messages are classified using rule-based multiple classifiers. The implementation of real-time processing of unstructured social media messages is a major flaw of the new survey. The solutions mentioned above do not address matching the attributes of these products and services with any other sets of messages.

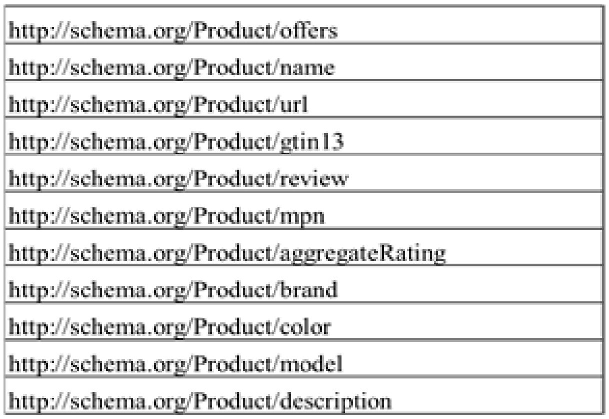
This paper presents a platform to extract product attributes from extract microblogging messages. Our architecture contains Natural Language Processing (NLP) algorithms that use low-latency processing to pull social media posts from social media platforms, such as Twitter and Facebook. Text analysis technologies such as part-of-speech, labeling, dictionary list(gazetteers), tokenizers, n-grams, and string similarity techniques are employed to process text. Logistic regression was utilized for classification, while Named-Entity Recognition (NER) was performed with CRF. We can attain excellent accuracy because of this new mix of technologies. For example, in our experiments, semantic extraction of product features on Twitter yielded an accuracy of more than 82 percent.

The remainder of the paper is organized as follows: Sec. II presents the recommended design; Sec. III presents the results of the study. Section III describes the performance evaluation process that was carried out utilizing a collection of real-world social media messages and evaluation criteria. Finally, in Section IV, we give some concluding thoughts.

II. PROPOSED DESIGN

A. System overview

Table I: Subset of references of RDF Product .



http://schema.org/Product/offers
http://schema.org/Product/name
http://schema.org/Product/url
http://schema.org/Product/gtin13
http://schema.org/Product/review
http://schema.org/Product/mpn
http://schema.org/Product/aggregateRating
http://schema.org/Product/brand
http://schema.org/Product/color
http://schema.org/Product/model
http://schema.org/Product/description

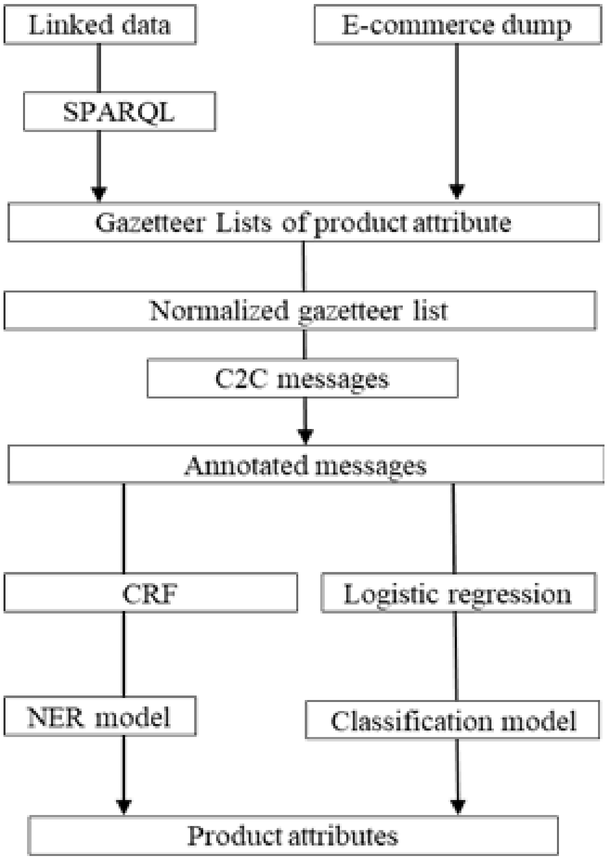


Figure 1: Information extraction process.

First, the text-analysis engine retrieves the product characteristics from social media messages that are not written down. The purpose of this research is to examine Twitter microblog messages precisely. A good example is when the information extraction module has finished processing the tweet “**Looking for LK88 PULSE oximeter for urgent need**". Following this, the information extraction module will generate various product attributes such as the brand of **PULSE**, the model of **LK88**, the product group of **health**, and the product is an **oximeter**. By the end of this process, the raw unstructured message has been structured, and later modules can now process it. Product group and status are identified using logistic regression classification, while brand and product semantics are derived using the CRF model.

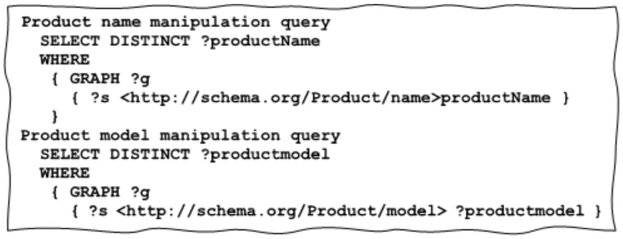
Table II: Steps In the Normalization of Product Brand’s Name

No	Steps to remove noisy brand names
1	The total number of brand names collected was around 13000
2	Brand names that have been normalized by the addition of other characters
3	Only alpha and alpha numeric brands were considered.
4	Removed extremely long and complicated brand names
5	With the assistance of World Net, I was able to remove all dictionary words.
6	Also deleted were stop words, but a few well-known brands were included, one of which was even in the dictionary.
7	It is permitted to use up to four word brands.
8	If the brand is a multi-word brand, then dictionary words and stop words are permitted.
9	The whole brand list was sorted in descending order based on the number of words and length.

B. The Extraction of Product Attributes

A list of text strings that are useful for named entity recognition (NER) is called a Gazetteer list. To avoid ambiguity, we created a gazetteer list for each attribute, including brand, product, and model. This investigation discovered many sets of data relevant to product information, such as the specifics recorded on Wikipedia, online product listings, and Amazon product reviews. The data sets used in this experiment are semi-structured. Based on resource and processing time limits, 250 distinct product-related features were discovered from Wikipedia. we used an Amazon product data set with certain semi-structural elements (He McAuley, 2016). A dataset in this collection includes product reviews, rating, usefulness, usefulness votes, and metadata (such as metaphors, group data, cost, brand name, and image characteristics) as well as link activity (number of reviews, visited and bought graphs) created between May 1996 and October 2018. (Ni, Li, and McAuley, 2019).

The total number of product entries in this Amazon dataset is roughly 233 million. We further took advantage of web data commons.org's gold standard for product feature data extraction and built the resulting feature data set into our product information (Petrovski et al., 2016). The data set consists of 232 million rows.



```
Product name manipulation query
  SELECT DISTINCT ?productName
  WHERE
   { GRAPH ?g
     { ?s <http://schema.org/Product/name>productName }
   }
Product model manipulation query
  SELECT DISTINCT ?productmodel
  WHERE
   { GRAPH ?g
     { ?s <http://schema.org/Product/model> ?productmodel }
```

Figure 2: Product attribute manipulation SPARQL queries on RDF based Linked data / Knowledgebase.

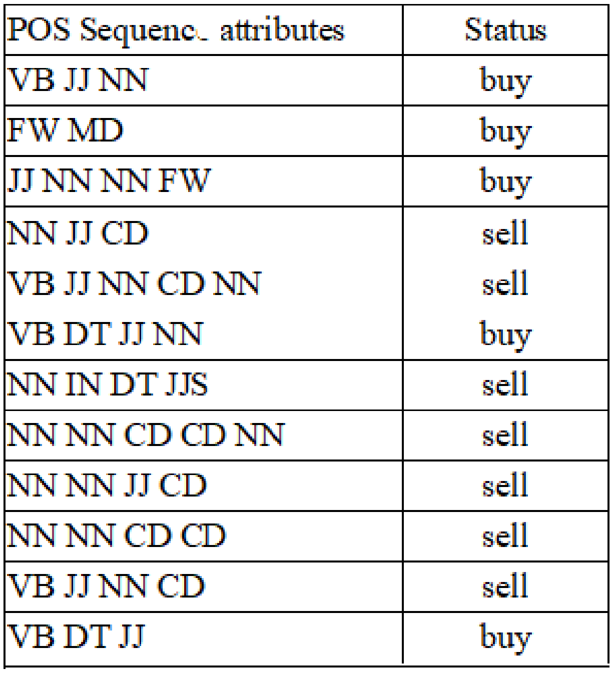
There are many product characteristics between different items. 36 attributes were contained in a headphone, 32 attributes were contained in a phone, and 76 attributes were contained in a television. To prevent ambiguity, the solution was limited to semantics like product name, product group, brand, model, and a few others in this work. An extensive gazetteer of various product attributes is produced from our research, which utilizes the Resource Description Framework (RDF) data store. The characteristics of the product are described in the table above labeled TABLE I.

In this experiment, the SPARQL protocol (protocol for RDF query language) was implemented to handle linked data/knowledge base data. In Figure 2 you can see a few of the SPARQL queries. The data for this research dealt specifically with brand and product categories. For every 500 products, 338 unique properties could be specified. To simplify processing, all product attributes were formatted as JSON objects. The data set used to train our machine learning model to recognize named entities is labeled with a gazetteer list, which includes all of the data that was previously identified as relevant product attributes.

Two classification problems with regards to product category and commercial intent are at the same time being investigated (i.e., buy or sell intension). Each message should be assigned to a product group and the microblogging message should be assigned to a commercial intent of either sell, buy, or neither. The classification of the product category was based on a separate data set obtained from Amazon (Ni et al., 2019). As a result, the data was of lower quality due to it containing duplicated entries, entries written in mixed upper and lower case, long words, and an abundance of numbers, prefixes, suffixes, and symbols.

Accordingly, the normalization techniques listed in TABLE II are Applied. In the first step, words that were shorter than four letters were excluded. Once the stop words were removed, the words that were found in the article were removed using the Python NLTK library. Another point is that a significant number of brand names had fewer than five words and so were removed from the list. Then, numbers and symbols were used to weed out the possible entries. Named entities which are between one and four words are used in this research to generate multiple files, and these are generated depending on the number of words. Separate files were generated for each word count, for example, for entries such as maco, mage, and malo, as well as entries such as mosey life, mount pros, and music hallo. Brand names for which the single-word list was compiled are screened to see if they are dictionary words or not. To test this, the online database WordNet (Miller, 1995) was utilized because it enables a user to find all permutations of a dictionary word. The word was removed from the list if it is a dictionary word. Nonetheless, numerous popular brands, including Apple and Orange, were added. We also kept the words themselves, such as microelectronics.

Table III: Attributes of the Part of Speech Sequence.



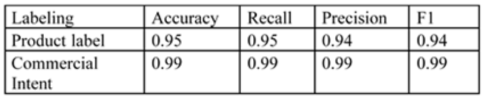
POS Sequenc_ attributes	Status
VB JJ NN	buy
FW MD	buy
JJ NN NN FW	buy
NN JJ CD	sell
VB JJ NN CD NN	sell
VB DT JJ NN	buy
NN IN DT JJS	sell
NN NN CD CD NN	sell
NN NN JJ CD	sell
NN NN CD CD	sell
VB JJ NN CD	sell
VB DT JJ	buy

In any case, whether they are words or not, entries with more than one word get the same weight. We found that grouping our list according to character count yielded the best results. At long last, we merged all three lists to produce a single list according to the word size and the length of each list. This led to a database of nearly 15,000 entries that serves as a brand gazetteer. Some components are critical to the C2C business model, as several goods have significant value for the model. A fifth dataset was necessary to discover which entity the data referred to. This included identifying a brand, a model, a product name, a product group, and whether a transaction was commercial or non-commercial. These gazetteers, however, were only intended to distinguish the manufacturer, model, and product name. We developed a real-time social media stream to get responses from social media sources like Twitter and additionally, the presence of product collection details on business-to-consumer and consumer-to-consumer websites such as Amazon. Complexity in dealing with these data sets results from the fact that the number of words in each element of the labeling dataset that is used to train the machine-learning model varies. Additionally, when doing things like training and labeling tasks in which we have no control over what words will be associated with a particular word in the Gazetteer, we can't expect that all of the words will match. GW Security Inc has a few differences compared to gwsecurity, such as the lack of a space between the words GW and security, as well as a missing word in the training set. Next is finding the intersection between two elements, for example, *smartworks consumer products* and *smartworks* itself. For two items to be included in our gazetteer, one of them must represent a different brand name. The other issue is the existence of multiple instances of the same type, which is typically known as duplicates. Thus, even after finding an entity, we must process the remainder of the text to determine whether or not there are any others like it.

An unsupervised graphical model based on conditional random fields (CRFs) (Song et al., 2019) One particular example of a linear-chain sequence, for instance, might be labeled a linear-chain sequence. Other positions on the sequence are also taken into consideration by CRF. As a result, the feature function will pay attention to the tokens situated next to the specific label. Eight to two: two-thirds of the training dataset is paired with one-third of the testing dataset. To assign each word a label, the corpus must first be tokenized. A BIO (beginning, in, out) tagged corpus was tokenized, and afterward, BIO (BEGINNING, IN, OUT) based tagging standards were put in place. One or more words were placed in a separate line for each label. A product that is missing the word in the label will be labeled as -O. It must be an introduction word at the beginning of the label, in which case it's designated -B. The expression will be formatted as follows: It was marked with a -I if the token it contained but not at the beginning of the string.

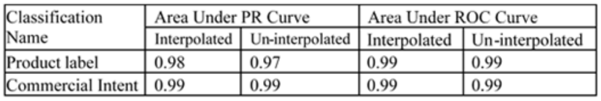
We classified the message based on its product group and business purpose during the classification step of the product-attribute extraction technique. Supervised machine learning methods rely on feature extraction to help the machine discover patterns and relationships in large data sets. For commercial intent, features such as having a URL, having product specifications, using specific terms, and using abbreviations were included in the TABLE III listing. With the Logistic Regression classifier, only the "Selling", "Buying", and "Neither" buckets were utilized in commercial intent classification.

TABLE V.CLASSIFICATION MODELS' ACCURACY.



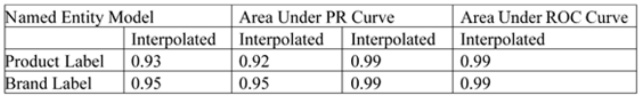
Labeling	Accuracy	Recall	Precision	F1
Product label	0.95	0.95	0.94	0.94
Commercial Intent	0.99	0.99	0.99	0.99

TABLE VI. PR AND ROC AREA MEASURES UNDER THE CLASSIFICATION CURVE.



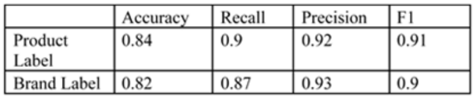
Classification Name	Area Under PR Curve		Area Under ROC Curve	
	Interpolated	Un-interpolated	Interpolated	Un-interpolated
Product label	0.98	0.97	0.99	0.99
Commercial Intent	0.99	0.99	0.99	0.99

TABLE VII. PR AND ROC AREA MEASURES UNDER THE NER CURVE.



Named Entity Model		Area Under PR Curve		Area Under ROC Curve
	Interpolated	Interpolated	Interpolated	Interpolated
Product Label	0.93	0.92	0.99	0.99
Brand Label	0.95	0.95	0.99	0.99

TABLE VIII. CRF MODEL ACCURACY MEASURES



	Accuracy	Recall	Precision	F1
Product Label	0.84	0.9	0.92	0.91
Brand Label	0.82	0.87	0.93	0.9

One of the leading classification algorithms that performed well with our dataset is the Logistic Regression algorithm. Because we did not have time to try other algorithms, Logistic Regression was used for the commercial intent classification. Both product type and commercial intent were taken into consideration to help classify this Logistic Regression-based classification problem (Hollerit et al., 2013). The usage of specific and unique POS sequence patterns is even more critical in commercial intent classification because commercial messages generally have few specific and unique POS sequence patterns, as listed in Table IV. Selling and buying messages only become a consideration when a future match occurs.

III. ANALYSIS OF PERFORMANCE

The suggested platform's performance was evaluated using a proof of concept implementation. The dataset had one million messages sampled from Twitter, which were broadcast at varying speeds, such as ten thousand messages per second, or one million messages per second. Our experimental cluster was built with a mix of compute and storage resources. A few PCs equipped with Intel Core i7 processors were employed to create the information extraction model. Nodes were equipped with 16 GB of RAM. With a heap size of 12 GB, the Lingpipe (Baldwin Carpenter, 2003) library was used.

The information extraction module was built to extract data on five separate entities: the product type, the brand, the

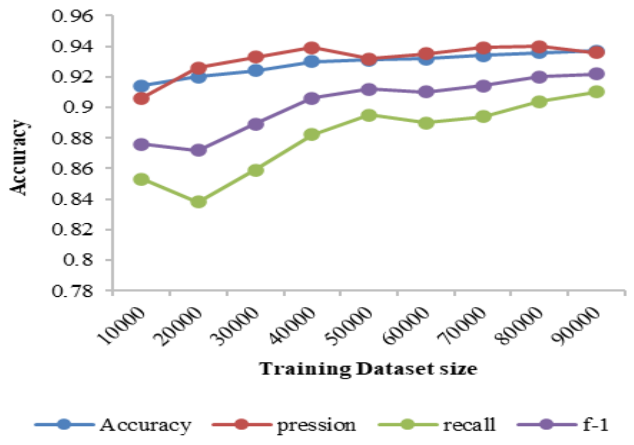


Figure 3: Accuracy of product group models in relation to the size of the training set.

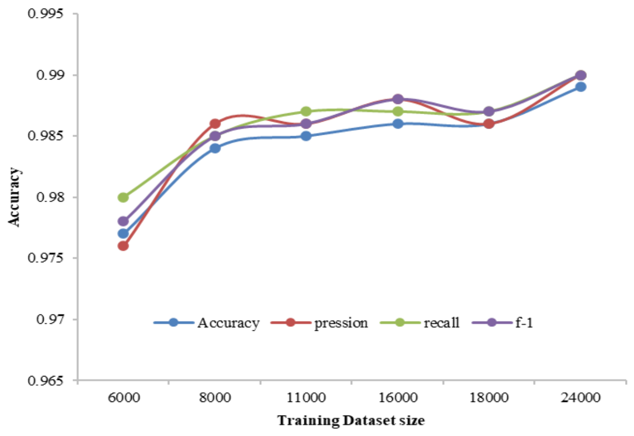


Figure 4: Accuracy of the selling status.

model, the selling status, and the product group. To begin, communications were classified by product category, including electronics, telephones, and music. A second classifier classified the signals as selling, purchasing, or neither. k-fold cross-validation was done to confirm that the results accurately represented the entire data set. Figures 3 and 4 show how accuracy grows over time as the dataset size increases. Furthermore, for both product group and commercial intent categorization, accuracy was attained. Table V lists the accuracy measures. The suggested classifier attained a classification accuracy of 0.95 for product groupings and 0.99 for commercial intent. In Tables VI and VII, ROC curves show When the model's threshold for recognizing a positive is modified, the recall-against-precision relationship modifications. As can be seen, the model's accuracy is verified by its closeness to the one region measuring of Precision-Recall (PR) and Receiver Operating Characteristic (ROC) under curve values. To finish identifying the remaining product qualities, the CRF was employed. These models both had an accuracy of 0.8, as shown in Table VIII.

IV. Conclusion

This research establishes a platform for matching microblogging posts suggesting buy and sell intentions in consumer-to-consumer e-commerce. To begin, a collection of algorithms based on CRF and NER were described for extracting product semantics and buy/sell intentions from unstructured social media communications. The tested platform's accuracy in classifying and matching product attributes, as well as its high throughput and low latency, was proved in a straightforward, low-cost hardware configuration that utilized Twitter data. We show the concept using microblogging messages, but it may easily be applied to other types of social media content that do not have a standard structure. Along with expanding the solution's support for additional products, services, and a dynamic number of attributes, we intend to enhance the solution in the future to enable multipart products and message matching based on user preferences.

References

BBaldwin, B., Carpenter, B. (2003). LingPipe. *Available from World Wide Web: Http://Alias-i. Com/Lingpipe.*

Bing, L., Wong, T.-L., Lam, W. (2016). Unsupervised extraction of popular product attributes from e-commerce websites by considering customer reviews. *ACM Transactions on Internet Technology (TOIT), 16*(2), 12.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Dan, C. (2014). Consumer-To-Consumer (C2C) Electronic Commerce: The Recent Picture. *International Journal of Networks and Communications, 4*(2), 29–32.

Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K. (2015). Analysis of Named Entity Recognition and Linking for Tweets. *Information Processing Management, 51*(2), 32–49. https://doi.org/10.1016/j.ipm.2014.10.006

Han, E.-H. (Sam), Karypis, G., Kumar, V. (2001). Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. In D. Cheung, G. J. Williams, Q. Li (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 53–65). Springer. https://doi.org/10.1007/3-540-45357-1_9

He, R., McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *Proceedings of the 25th International Conference on World Wide Web*, 507–517.

Hollerit, B., Kröll, M., Strohmaier, M. (2013). Towards Linking Buyers and Sellers: Detecting Commercial Intent on Twitter. *Proceedings of the 22Nd Inter-*

national Conference on World Wide Web, 629–632. https://doi.org/10.1145/2487788.2488009

Jenhani, F., Gouider, M. S., Said, L. B. (2018). Social Stream Clustering to Improve Events Extraction. In I. Czarnowski, R. J. Howlett, L. C. Jain (Eds.), *Intelligent Decision Technologies 2017* (Vol. 73, pp. 319–329). Springer International Publishing. https://doi.org/10.1007/978-3-319-59424-8$_{3}$0

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Ni, J., Li, J., McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 188–197.

Nozza, D., Ristagno, F., Palmonari, M., Fersini, E., Manchanda, P., Messina, E. (2017). TWINE: A real-time system for TWeet analysis via Information Extraction. *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* 25–28. https://doi.org/10.18653/v1/E17-3007

Petrovski, P., Primpeli, A., Meusel, R., Bizer, C. (2016). The WDC gold standards for product feature extraction and product matching. *International Conference on Electronic Commerce and Web Technologies,* 73–86.

Raju, S., Pingali, P., Varma, V. (2009). An unsupervised approach to product attribute extraction. *European Conference on Information Retrieval,* 796–800.

Song, S., Zhang, N., Huang, H. (2019). Named entity recognition based on conditional random fields. *Cluster Computing, 22*(3), 5195–5206.

Tarasconi, F., Farina, M., Mazzei, A., Bosca, A. (2017). The role of unstructured data in real-time disaster-related social media monitoring. *2017 IEEE International Conference on Big Data (Big Data),* 3769–3778. https://doi.org/10.1109/BigData.2017.8258377