

Exploring User-Generated Content to Detect Community Problems: The Ontological Model of ALLEGRO

Carlos Perrián-Pascual

*Universitat Politècnica de València, Paranimf 1, 46730 Gandia (Valencia), Spain
joepas3@upv.es*

Keywords: User-Generated Content, Problem Detection, Text Classification, Keyword Recognition, Ontology.

Abstract: Social-media services contribute to creating situation awareness, thus offering a snapshot of today's society. Citizens can use such communication channels to report problems concerning the quality of life of individuals and the well-being of the community in which they live. Therefore, we can develop applications that can analyse online user-generated data about a variety of problems from different topics (e.g. education, health, or politics, among many others) to reconstruct the state of society as interpreted by social-media users in the given community. In this context, the main objective of this paper is to describe the ontological model required for representing community problems affecting quality of life and well-being, and how this ontology supports the natural language processing and text-mining tasks of topic categorisation and keyword extraction. This ontological model can become a significant component in natural language understanding applications, particularly in those where machine-learning or neural-network models are enhanced with external knowledge to perform opinion mining.

1 INTRODUCTION

Social-media services (e.g. Twitter, Facebook etc.) have become a global phenomenon of communication, where users post content in the form of text, images, video, audio or a combination of them to convey their opinions, report facts, or show situations of interest. A current line of research related to these tools consists in crowdsensing, i.e. the analysis and interpretation of the massive amount of user-generated content (UGC) posted daily in these communication channels. In this context, this paper results from the ongoing research conducted in the ALLEGRO project (Adaptive multi-domain social-media sensing framework), a general-purpose multi-modal system (i.e. text, audio, and image) for the development of applications that can accurately reconstruct the state of society as interpreted by the collective intelligence of social-media users. In other words, in the framework of social-media analytics, we intend to sense UGC to construct models that can detect community problems, thus considering users as witnesses of a given society.

The remainder of this paper is organised as follows. Section 2 describes relevant works for the

context of this study. Section 3 provides an overview of the ALLEGRO system and gives a detailed account of the DIAPASON module, particularly the ontological model and its role in the pipeline of natural language processing (NLP) and text-mining tasks involved in topic categorisation and keyword recognition. Finally, Section 4 presents some conclusions.

2 RELATED WORK

2.1 Twitter for Smart Societies

In the last few years, the concept of Smart City has been moving towards the notion of Smart Society (Valkenburg et al., 2016), where citizens should be engaged to actively participate in creating a higher quality of life for themselves and others. To this end, citizens should be provided with a space to participate and be involved "if an open, multipurpose democratised platform is applied in the public domain, data can empower people to become active producers of societal value" (Valkenburg et al., 2016, p. 91).

In this context, social media can be regarded as a shared platform for participating citizens so that people become truly empowered in smart societies. For this reason, the latest research efforts focus on extracting knowledge from social media to contribute to developing smart-city systems, where taking Twitter as a sensor has gained increasing interest because of the real-time nature of its data. In this regard, two controversial issues could seemingly distort research results: restrictions on data acquisition and the veracity of information. However, such issues do not undermine the adequacy of Twitter as an information-providing platform for smart-city applications. On the one hand, the Twitter Streaming API is only able to return up to 1% of all content published at a given time, but Ayora et al. (2018) empirically demonstrated that neither the lack of completeness nor the latency of Twitter data results in flawed data that could lead to wrong conclusions. On the other hand, Doran et al. (2016) presented the reasons for relying on the collective information conveyed by a stream of UGC:

(a) users lose credibility within their social networks when they continuously share posts that are unlikely to be authentic and truthful, and

(b) false information provided by a few is unlikely to lead to misleading inferences because truthful information is usually shared by an overwhelming majority.

Two projects merit our attention in this framework of Smart Society, particularly in developing general-purpose Twitter-based crowdsensing systems. For example, TwitterSensing (Costa et al., 2018) detects and classifies events of interest (e.g. accidents, floods, traffic jams, etc.) not only to enhance the quality of wireless sensor networks but also to detect the areas where new sensors are required. In this case, information from events that are currently happening is extracted when tweets are converted into vector representations using term frequency and inverse document frequency (TF-IDF), which are then fed into a Multinomial Naïve Bayes classifier.

Adikari and Alahakoon (2021) proposed an AI-based system to monitor the 'emotional pulse' of the city by analysing the emotions collectively expressed by citizens through data from social media and online discussion forums. The system carries out three main tasks. First, primary emotions (i.e. anger, anticipation, disgust, fear, joy, sad, surprise, and trust) are extracted based on a crowdsourced lexicon for emotion mining (Mohammad & Turney, 2018); as a result, an emotion profile is created for each

tweet. Second, emotion transitions are modelled using Markov models. Finally, toxic comments, which indicate a higher level of negativity than basic negative emotions, are detected with a deep-learning multi-label classifier, which employs layers of word embedding, bidirectional recurrent neural networks and convolutional neural networks.

On the other hand, most of the latest studies that integrate social media into a smart-city model focus on specific domains, such as traffic (Pandhare & Shah, 2017; Lau, 2017; Salas et al., 2017), healthcare (Alotaibi et al., 2020) or security (Saura et al., 2021). For example, Pandhare and Shah (2017) proposed a system that detects tweets related to traffic and accidents. After filtering out stopwords, they determined the importance of the tokens in a tweet through TF-IDF and then employed logistic regression and SVM as binary classifiers (i.e. traffic and non-traffic tweets).

Lau (2017) extracted useful driving navigation information (e.g. road accidents, traffic jams, etc.) from social media (i.e. Twitter and Sina Weibo) to enhance the effectiveness of Intelligent Transportation Systems, which provide drivers with real-time navigation information. First, he employed a topic model-based method (Latent Dirichlet Allocation) to learn concepts about traffic events from an unlabeled corpus of UGC (i.e. message filtering). Second, he applied an ensemble-based classification method to detect traffic-related events automatically (i.e. event identification). In particular, the ensemble classifier relied on a weighted voting scheme with three base classifiers, i.e. support vector machines (SVM), Naïve Bayes, and K-Nearest Neighbour.

Salas et al. (2017) proposed a framework to analyse real-time traffic incidents using Twitter data, where the main steps are as follows. First, tweets are tokenised, and stopwords and special characters are removed. Second, tweets are classified into traffic-related and non-traffic-related, and traffic-related tweets are in turn classified into different event categories: roadworks, accidents, weather, and social events; SVM is used as the classifier. Third, the tweet location is extracted using named-entity recognition and entity disambiguation based on Wikipedia. Fourth, the strength of positive or negative sentiment (ranging from -5 to 5) is predicted for each tweet. Finally, the level of stress or relaxation for each tweet (ranging from -5 to 5) is also determined; therefore, when the user is complaining, the level of stress in the text is high.

Alotaibi et al. (2020) developed a big-data analytics tool to detect symptoms and diseases using

Twitter data in Arabic and thus report the top diseases in the Kingdom of Saudi Arabia. To this end, they manually annotated a sample of tweets as related, i.e. reflecting a health concern, or, otherwise, unrelated. In turn, health-related tweets were further labelled as (a) messages about actual cases of sickness, suffering, and medication, or (b) posts creating awareness about health problems. The highest accuracy was obtained by classifying tweets with Naïve Bayes based on trigrams.

Saura et al. (2021) applied sentiment analysis to a dataset of about 750,000 tweets to detect positive, negative and neutral tweets, where linear support vector classifiers and logistic regression obtained the best results in accuracy. Then, they employed Latent Dirichlet Allocation to divide the sample into security-related topics, automatically grouped into keywords according to their frequency.

2.2 Citizen-Reporting Tools for Smart Cities

In the context of Citizen Relationship Management, where the government tends to see citizens as customers (Kopackova et al., 2019), a wide range of smartphone applications has been developed to report everyday non-emergency problems about urban infrastructure and services (e.g. air pollution, traffic congestions, potholes in roads, and broken lights, among many others). Such problematic issues can decrease citizens' quality of life within their communities. Therefore, these participatory tools are beneficial not only for citizens, who can reach the local government to fix problems in their neighbourhood, but also for the local government, who can get information about what citizens want and need (Kopackova et al., 2019), where the economic impact for the latter also deserves to be highlighted, who 'might find out about issues sooner, fix them and spend less on paid inspectors who would travel around the urban areas and look for potholes, garbage and similar' (Lendák, 2016, p. 358). These crowdsensing applications, e.g. FixMyStreet,¹ PublicStuff,² SeeClickFix,³ Novoville⁴ and Improve My City,⁵ among others, usually consist of a mobile application that citizens use for sensing and a website that shows the sensed events to the administration in near real-time. Indeed, most

of these applications share similar features in the data and sensing layers of their architecture. In the data layer, such applications deal with structured data submitted to administrators, who will directly manipulate them to access information. In the sensing layer, heterogeneous data in the form of text and/or images are contributed by the user when a given event occurs; in other words, when citizens recognise the event (e.g. car accident), they perform the sensing task (e.g. taking a picture).

3 ALLEGRO

3.1 Architecture

ALLEGRO consists of two modules (i.e. Data Analysis and Data Fusion), which employ a multi-modal data repository and a knowledge base. The Data Analysis module is comprised of a dedicated component for each type of data to be analysed in UGC, i.e. DIAPASON (text analysis), ADAGIO (image analysis), and SOUND (audio analysis). In this regard, microtext analysis is viewed as the initial process that provides the context of the problem described in each message so that an instantiation of a given problem schema can be returned. In case that messages go with embedded audio and/or image content, this schema can be supplemented with context data from audio analysis and/or image analysis, which are concurrently executed, both to verify or rebut event-related information detected in the text or to complete missing information in the knowledge schema. Then, augmented knowledge schemas produced in this module are combined in the Data Fusion module, where the quality of aggregated data is enhanced by rejecting irrelevant information, minimising redundancy, resolving inconsistencies, and completing missing information. In this context, any of the components in ALLEGRO (i.e. DIAPASON, ADAGIO, and SOUND) relies not only on the same ontology to model the knowledge extracted from its corresponding subtype of UGC item (i.e. text, image, and voice message, respectively) but also on the same formalism (i.e. problem schema) to represent the most distinctive aspects of problem types. Both issues are described and illustrated with DIAPASON in Section 3.2.

ALLEGRO is a crowdsensing system developed in the framework of Smart Society. The contributions of ALLEGRO with respect to the smart-city systems that extract knowledge from social media (Section 2.1) and citizen-reporting tools

¹ <https://www.fixmystreet.com>

² <http://www.publicstuff.com>

³ <http://en.seeclickfix.com>

⁴ <http://www.novoville.com>

⁵ <http://www.improve-my-city.com>

for smart cities (Section 2.2) can be found in three main aspects:

- It encompasses a wider variety of social and physical problem types in society from the perspective of the citizens who live in a given community.
- It performs a deeper analysis of collected unstructured data through topic categorisation and keyword recognition to make predictions from automatically extracted information.
- It aims to perform multi-modal data fusion from the knowledge derived from its core modules—i.e. DIAPASON (text), SOUND (audio), and ADAGIO (image).

3.2 DIAPASON

3.2.1 Ontology

One of the primary components in ALLEGRO is DIAPASON (unified hybrid Approach to microtext Analysis in Social-media crowdsensing), a proof-of-concept workbench where English and Spanish short texts from social media can be analysed by integrating natural language processing, machine and deep learning, and knowledge engineering techniques. Indeed, DIAPASON adopts a hybrid approach to artificial intelligence, where a knowledge-based system relies on neural-network models for text classification, thus combining human intelligence with machine intelligence.

Organising knowledge is a critical issue in smart cities. In this regard, ontologies play a critical role, being often defined in knowledge engineering and artificial intelligence as "a specification of a representational vocabulary for a shared domain of discourse — definitions of classes, relations, functions, and other objects" (Gruber, 1993, p. 199). The DIAPASON ontology consists of four levels of classes (i.e. problem realm, problem dimension, problem domain, and problem type) modelled in a single hierarchy, being PROBLEM the superclass at the top. Thus, knowledge is organised around two primary realms: SOCIAL and PHYSICAL. In turn, this upper level is structured into six dimensions, where the LIVING, ECONOMY and GOVERNANCE dimensions pertain to the social realm, and the MOBILITY, INFRASTRUCTURE and ENVIRONMENT dimensions to the physical realm. The modelling of this level results from smart-city frameworks such as Giffinger et al. (2007), Govada et al. (2017), and Appio et al. (2019). At the middle level, we organised each dimension in domains on which citizens can show an attitude of disapproval towards

some specific aspect of the community. For example, the class PHYSICAL subsumes the class ENVIRONMENT, which represents the problems about the environmental dimension, which in turn subsumes the class ECOLOGICALHAZARD, which represents the problems about the ecological-hazard domain. The levels of problem realms, problem dimensions, and problem domains in the DIAPASON ontology are currently fully developed. Finally, the lower level describes types of community problems that can affect some, most or all citizens. For example, the classes MARINELITTER and MARINESPILL, which are subsumed by the class ECOLOGICALHAZARD, address the environmental quality of urban and urbanised beaches in the framework of Smart Cities (Ariza et al., 2010; Sardá et al., 2014). The level of specific problem types in the DIAPASON ontology is currently under development because of the numerous elements that can be found. Each specific problem type is modelled as a distinct class subsumed by one or more classes at the domain level.

It should be noted that we formalise each problem type, which pertains to one or more domains, using a language-independent problem schema. For example, the problem schemas of MARINELITTER and MARINESPILL are presented in (1) and (2), respectively.

- (1) ((plastic-114592610 | bag-102773037 | bottle-102876657 | glass-103438257 | cap-102954938 | lid-103661340 | butt-102927399 | can-102946921 | dirty-300419289) & (sand-115019030 | sea-109426788))
- (2) (((oil-114980579 | spill-115049594 | tar-114911704) | (dead-300095280 & fish-102512053)) & (sand-115019030 | sea-109426788))

Formally, our problem schemas consist of five types of elements (i.e. concepts, named entities, functions, operators, and expressions):

- Concepts are WordNet synsets (Fellbaum, 1998), i.e. sets of synonymous words. WordNet is a lexical database that organises nouns, verbs, adjectives, and adverbs into synsets, where each synset represents a distinct concept connected to other synsets through lexical and semantic relations. In (1) and (2), we introduce each synset with an English word (e.g. plastic-114592610) to facilitate the readability of the

problem-type representation; however, such words play no role during the processing of problem schemas.

- Named entities are real-world entities (e.g. individuals, locations, organisations, etc.) denoted by proper names. From an ontological perspective, named entities in problem schemas can take the form of instances (e.g. \$Mediterranean_Sea) or instance categories (e.g. \$sea), which serve to cluster a set of instances (e.g. *Adriatic Sea*, *Black Sea*, *Mediterranean Sea*, etc.).
- Functions represent lexical, syntactic, and semantic patterns that can be implemented in various linguistic realisations to express pragmatic meaning. To illustrate, DISAPPOINTED is linguistically projected to constructions such as *I was hoping for...*, *I'm sorry to hear...*, and *it's a real pity...*, among many others.
- Operators are categorised into two groups. On the one hand, conceptual operators (i.e. modifiers and negation) act on the concepts of a proposition to give greater semantic specificity to the description of the problem type. Modifiers aim to present a particular quality, entity, event or situation at a higher (M) or lower (P) quantity, degree or intensity than expected for the state of affairs being described. In contrast, negation (N) refers to any construct that introduces the lack of a given quality, entity, event or situation. Conceptual operators can be mapped to linguistic realisations. On the other hand, logical operators, i.e. conjunction (&), inclusive disjunction (|) and exclusive disjunction (^), can connect two or more elements of the same kind.
- Expressions are constructs that include concepts, named entities, functions, or other expressions, providing that such elements are connected through logical operators. Round brackets determine the scope of expressions.

3.2.2 Text Processing

The goal of DIAPASON is to discover the specific situations described by UGC that can instantiate one or more problem types. To this end, a pipeline of NLP and text-mining tasks is performed:

a) Pre-processing texts. This task, which aims to have clean texts from the input, is especially relevant to UGC since the language used in social media is characterised by departing from the commonly

accepted standard for written text. In the case of tweets, emojis are converted into lexical units, hashtags are segmented into tokens, and references and URL links are automatically removed, among other standardisation methods.

b) Processing texts. First, each text is tokenised, and all the tokens are lemmatised. Second, named entities are recognised and linked to entity types, which become part of the tokenised input. Third, nouns, verbs and adjectives are conceptualised by identifying their corresponding WordNet synsets, and negation cues and modifiers are projected as conceptual operators. Such conceptualisation is performed with unsupervised word sense disambiguation based on synset embeddings. Fourth, functions were detected by adopting a rule-based approach grounded on the lexical, syntactic and semantic features of the input. Therefore, elements such as named entities, concepts and functions in problem schemas play a critical role in this stage.

c) Classifying texts. Identifying which texts describe which problems is regarded as a text-categorisation task. To this end, a corpus of labelled instances is created from the knowledge stored in problem schemas so that a two-dimensional Convolutional Neural Network model is constructed from pre-defined embeddings in LessLex (Colla et al., 2020) to make predictions on new UGC. Therefore, the embeddings linked to the WordNet synsets of named entities, concepts and functions make up the foundation of this stage.

d) Recognising keywords from texts. Once the multi-domain categorisation of a single text has been performed, the system represents the semantics of each token in the text and each synset in the relevant problem schemas as embeddings computed from the language model. As such embeddings are located in the same high-dimensional vector space, each word embedding is compared with each synset vector based on cosine similarity, assuming that the closer an individual word embedding is to a given synset vector, the more likely the synset becomes a descriptor. Finally, keywords are obtained through the words linked to the selected synsets. Therefore, keyword recognition is primarily performed through the embedding-based similarity between UGC and problem schemas.

As a way of summary, Figure 1 illustrates the process where NLP and text-mining tasks are involved.

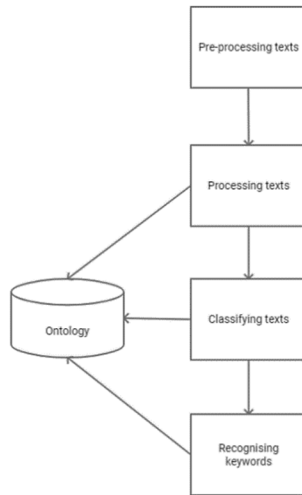


Figure 1: Text processing in DIAPASON.

To illustrate, suppose that we intend to explore a collection of tweets to discover water-related environmental problems in Australia. For example, the tweet in (2) could pertain to such a corpus.

- (2) #Cigarette butts, the most common source of #litter in Victorian waterways, can take up to 12 months to break down in freshwater and up to five years to break down in seawater. yarraandbay.vic.gov.au/issues/litter

After text processing, DIAPASON obtains the information shown in Table 1 from the metadata and core data of the UGC unit whose text message is (2).⁶

Table 1: Twitter information extraction: an example.

Tweet ID	1022270465144971264
Timestamp	2018-07-26T02:00:00
Language	en
Problem	MARINELITTER
Named entity	Victorian waterways
Synsets [keyword]	102927399 [butt] 114858292 [litter] 115008847 [seawater]

According to Beliga et al. (2015), keyword-selection methods can be divided into two categories: keywords can be selected from a controlled vocabulary of terms (i.e. keyword assignment) or directly from the source document (i.e. keyword extraction). In DIAPASON, keyword recognition adopts the former approach, as keywords

are derived from problem schemas. However, not all words related to the synsets in problem schemas are instantiated as keywords, but only those with a significant semantic association with some lexical unit in the tweet determined by distributional semantics.

4 CONCLUSIONS

Being developed in the framework of Smart Society, ALLEGRO is an ongoing project based on the processing of UGC contributed by social sensors. As UGC can be comprised of different types of data, e.g. text, audio, image, and video, the Data Analysis module in ALLEGRO includes a dedicated component for processing each of these types of data. In this regard, DIAPASON, which has been devised for text analysis, is provided with an ontology that is being constructed to cover a wide variety of community problems from the perspective of social-media users. This ontology is organised into four levels of topic specificity, where the lower level includes classes that represent specific types of community problems. In turn, problem schemas are assigned to such specific types, as they serve as conceptual representations of their semantics. Indeed, problem schemas play an active role in the multi-domain categorisation of UGC, as well as in keyword extraction.

ACKNOWLEDGEMENTS

This article was supported under grant PID2020-112827GB-I00 funded by MCIN/AEI/10.13039/501100011033, and under grant number 101017861 [project SMARTLAGOON] by the European Union's Horizon 2020 research and innovation program.

REFERENCES

- Adikari, A., & Alahakoon, D. (2021). Understanding citizens' emotional pulse in a smart city using artificial intelligence. *IEEE Transactions on Industrial Informatics*, 17(4), 2743-2751.
- Alotaibi, S., Mehmood, R., Katib, I., Rana, O., & Albeshri, A. (2020). Sehaa: A big data analytics tool for healthcare symptoms and diseases detection using Twitter, Apache Spark, and machine learning. *Applied Sciences*, 10(4), 1-29.

⁶ https://mobile.twitter.com/epa_victoria/status/1022270465144971264

- Ariza, E., Jimenez, J. A., Sarda, R., Villares, M., Pinto, J., Fraguell, R., Roca, E., Marti, C., Valdemoro, H., Ballester, R., & Fluvia, M. (2010). Proposal for an integral quality index for urban and urbanized beaches. *Environmental Management*, 45, 998-1013.
- Ayora, V., Horita, F., & Kamienski, C. (2018). Social networks as real-time data distribution platforms for smart cities. In *Proceedings of the 10th Latin America Networking Conference* (pp. 2-9). Association for Computing Machinery.
- Beliga, S., Mestrovic, A., & Martincic-Ipsic, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences*, 39(1), 1-20.
- Colla, D., Mensa, E., & Radicioni, D. P. (2020). LessLex: Linking multilingual embeddings to sense representations of lexical items. *Computational Linguistics*, 46(2), 289-333.
- Costa, D. G., Duran-Faundez, C., Andrade, D. C., Rocha-Junior, J. B., & Just Peixoto, J. P. (2018). Twittersensing: An event-based approach for wireless sensor networks optimization exploiting social media in smart city applications. *Sensors*, 18(4), 1-30.
- Doran, D., Severin, K., Gokhale, S., & Dagnino, A. (2016). Social media enabled human sensing for smart cities. *AI Communications*, 29, 57-75.
- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press.
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), 199-220.
- Kopackova, H., Komarkova, J., & Jech, J. (2019). Technology helping citizens to express their needs and improve their neighborhood. In *Proceedings of the 2019 International Conference on Information and Digital Technologies* (pp. 229-236). IEEE.
- Lau, R. Y. (2017). Toward a social sensor based framework for intelligent transportation. In *Proceedings of the 18th International Symposium on a World of Wireless, Mobile and Multimedia Networks* (pp. 1-6). IEEE.
- Lendák, I. (2016). Mobile crowd-sensing in the smart city. In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, & R. Purves (Eds.), *European handbook of crowdsourced geographic information* (pp. 353-369). Ubiquity Press.
- Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436-465.
- Pandhare, K. R., & Shah, M. A. (2017). Real time road traffic event detection using Twitter and spark. In *Proceedings of the 2017 International Conference on Inventive Communication and Computational Technologies* (pp. 445-449). IEEE.
- Salas, A., Georgakis, P., Nwagboso, C., Ammari, A., & Petalas, I. (2017). Traffic event detection framework using social media. In *Proceedings of the 2017 IEEE International Conference on Smart Grid and Smart Cities* (pp. 303-307). IEEE.
- Sardá, R., Ariza, E., Jiménez, J. A., Valdemoro, H., Villares, M., Roca, E., Pintó, J., Martí, C., Fraguell, R., Ballester, R., & Fluvía, M. (2014). El índice de calidad de playas (BQI). In R. Sardá, J. Pintó, & J. Francesc Valls (Eds.), *Hacia un nuevo modelo integral de gestión de playas* (pp. 105-122). Documenta Universitaria.
- Saura, J. R., Palacios-Marqués, D., & Ribeiro-Soriano, D. (2021). Using data mining techniques to explore security issues in smart living environments in Twitter. *Computer Communications*, 179, 285-295.
- Valkenburg, R., Den Ouden, E., & Schreurs, M. A. (2016). Designing a smart society: From smart cities to smart societies. In B. Salmelin (Ed.), *Open Innovation 2.0 yearbook 2016* (pp. 87-92). European Commission.