INTI JOURNAL | Vol.2023:71
eISSN:2600-7320

An In-Depth Analysis of Text Clustering Techniques for Identifying Potential Insurance Customers on Social Media: A Machine Learning Perspective

Liew Chun Kin[1], Goh Ching Pang[1*]

[1] Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia

[*]**Email:** gohcp@tarc.edu.my

Abstract

Social media has emerged as a transformative platform for the exchange and dissemination of information. Unlike conventional sources such as online news, social media often offers more real-time and current updates. Effectively harnessing the vast and diverse pool of unstructured data on these platforms requires the extraction of structured information. This research focuses on the development of a social media web crawler, coupled with the implementation of sophisticated algorithms like Web Content Mining, Noisy Text Filtering, Named Entity Extraction, Part-Of-Speech (POS) Tagging, and Text Clustering. The aggregated information will be utilized to train a machine learning model capable of discerning a customer's preferred insurance type—be it accident, health, car, or life insurance. The overarching objective is to provide insurance companies with a swift, precise, and cost-effective means of identifying potential customers within the realm of social media. The result shows that this new technique has successfully identify relevant topic based on the comments and recommend corresponding insurance to the user.

Keywords

LDA bag of words, LDA TF-IDF, insurance, machine learning

Introduction

The insurance industry has undergone a significant transformation in recent years, primarily driven by rapid advancements in information technology. This shift is particularly evident in how insurance companies now leverage data and social media to enhance their services and customer engagement. In an era where consumers increasingly seek transparency and avoid high-pressure sales tactics, understanding and utilizing the wealth of data available online becomes crucial. This is especially true in the context of social media, which has emerged as a novel medium for information exchange and sharing. Platforms like Twitter and Facebook epitomize this trend, forming virtual communities where individuals and organizations alike create, share, and exchange information and ideas.

According to recent reports, such as one from CNN1, a growing number of Americans now rely on the Internet, especially social media, for news, with three-quarters citing online sources like email updates or social media feeds for breaking news. This shift highlights the pivotal role

Submission: 14 November 2023; **Acceptance:** 6 December 2023

of social media in outpacing traditional sources, including online news, in providing timely information (Habib, 2014). The insurance sector, recognizing the potential of this vast data reservoir, is increasingly focusing on Information Extraction (IE) to make sense of the unstructured data prevalent on the web. IE has become a critical research domain, enabling the systematic utilization of vast amounts of unstructured distributed data (Hu and Xia, 2021; Akhmadeeva et al., 2016). IE systems are designed to analyze human language text to extract valuable information about events, entities, or relationships. This information is then structured into Knowledge-bases (KBs), which are repositories that store facts and relationships gleaned from text, making the data accessible and useful for both machines and humans (Dong et al., 2020). However, despite extensive research in IE, there has been limited focus on social media content. Tools like TwitIE, an open-source NLP pipeline designed for microblog text, have been developed (Bontcheva et al., 2013), but they often lack comprehensive capabilities like message filtering, named entity disambiguation, and relation/fact extraction, which are critical for the insurance industry to analyze and understand customer needs and market trends effectively.

To enhance these capabilities, machine learning algorithms, particularly those trained to extract various entities and relations from both structured and unstructured texts, are being employed (Sarker, 2021; Jones & Sah., 2023; Rawat et al., 2021). Nonetheless, the supervised training of these algorithms is resource-intensive and requires numerous labeled examples for each entity type and connection. To mitigate this, researchers are exploring semi-supervised learning methods that use limited labeled data alongside large volumes of unlabeled text. Despite their promise, these methods often face challenges in accuracy and consistency, particularly when controlling the learning process with a small number of initial labeled samples (Carlson et al., 2010). In summary, the insurance industry's evolution is a testament to the power of technology and data. By embracing the digital age, particularly the wealth of information available through social media, insurance companies are not only improving their understanding of consumer needs but also revolutionizing the way they connect with and serve their customers.

Methodology

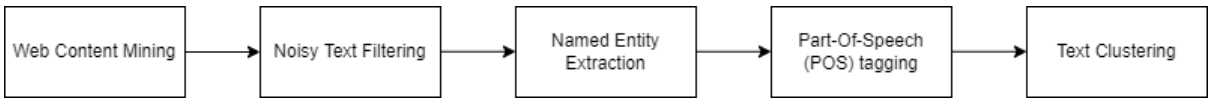

Figure 1. Flowchart of the research

The research is structured around two key components as shown in Figure 1: web content mining and text clustering. The initial phase involves the creation of a web crawler using Python, emphasizing text extraction from online documents. The text undergoes a crucial noisy text filtering process to eliminate irrelevant posts, followed by named entity extraction and part-of-speech tagging. In the second phase, attention shifts to text clustering, where unlabeled texts are grouped to create clusters exhibiting greater similarity within the same cluster. This structured approach aids in understanding patterns and relationships within the text corpus.

Snscrape emerges as the optimal social media scraping tool following rigorous independent testing. It excels in retrieving items through scraping user profiles, hashtags, and searches, supporting various services such as Facebook, Twitter, Instagram, and more. The tool's versatility

and proven performance make it a standout choice for effective social media scraping in the project. The scraped comments are shown in Figure 2. This research encompasses the scraping of three distinct comment types—food, car, and sport. Each category will be subjected to the scraping of 5000 comments for comprehensive analysis.

	Author_Name	Text
2	Cajun Silverado	My '06 Chevy Silverado 1500HD with 6.0 liter V8 is an awesome truck. I tow a 31' 5th wheel c
3	rick	I have owned 5 Silverado's since 1999, would not consider another truck, my new HD with the
4	Luvmytruck	I am a line driver for a local trucking company and when I get back to the yard I can't wait to g
5	Kevin VanAntwerp	We purchased this thruck to pull a 33 ft Americamp trailer (7500 lbs) and I'm pleased with the
6	Mark Bucher	This has been the best truck I've ever owned. It's been from Alaska to Florida towing a 17' tra
7	dyork920	my dad purchased this truck new in 01 we have had this truck for 10 years of camper haulin, c
8	BuiltChevyTough3	I bought this truck in February of last year. It had 142,000 miles on it. It now has 165,000 miles
9	Long-time owner	Have hauled 2 pallets (>5000#) of paving stones without a sweat. Towing package made it ea
10	Jeff	I recently purchased this truck and I know from the test drive it was going to be a great truck a

Figure 2. Scrapped comments from Twitter using snscrape (sample code snippet)

Text processing, leveraging natural language processing (NLP) and machine learning, analyzes and organizes unstructured text data. Text classification, a primary technique, categorizes and analyzes textual data based on content, including topic analysis, sentiment analysis, intent detection, and language classification. Text extraction identifies and extracts vital information using parameters like keywords, client names, and dates. Techniques like entity and keyword extraction, involving tokenization and stopwords removal, contribute to a refined text processing approach.

Linear Discriminant Analysis (LDA) (Kim et al., 2023) plays a pivotal role in machine learning's dimensionality reduction methods. Also known as Discriminant Function Analysis (DFA) or Normal Discriminant Analysis (NDA), LDA efficiently projects characteristics from higher-dimensional to lower-dimensional space, conserving resources. Unlike logistic regression limited to two-class problems, LDA is versatile, addressing challenges with more than two classes. It serves as both a dimensionality reduction method and a pre-processing stage for pattern categorization and modeling variations. LDA excels in distinguishing multiple classes amid numerous features, transforming 2-D or 3-D graphs into a 1-dimensional plane. This is demonstrated by efficiently categorizing two classes on a 2-D plane with an X-Y axis using a straight line, showcasing LDA's capability to optimize separation between classes.

Results and Discussion

Figures 3 to 5 present three illustrative examples of the test plan outcomes derived from running the testing procedure with distinct comments. In Figure 3, a comment discussing Bingsu and chicken rice in Singapore is categorized as food by the system. Subsequently, the system recommends health insurance, aligning with the identified food theme. Moving to Figure 4, the system adeptly recognizes a comment detailing a traffic accident on a Friday morning, appropriately categorizing it under the car topic. In response, the system recommends a comprehensive set of relevant insurances, including life, car, and accident coverage. Figure 5 illustrates a sports-related comment detailing a harrowing experience of almost suffocating during

climbing. The system accurately identifies the comment as pertaining to sports and recommends life and accident insurance tailored to the user's situation. These examples showcase the system's proficiency in categorizing diverse comments and providing relevant insurance recommendations based on the identified themes.

Figure 3. Test case on food comment

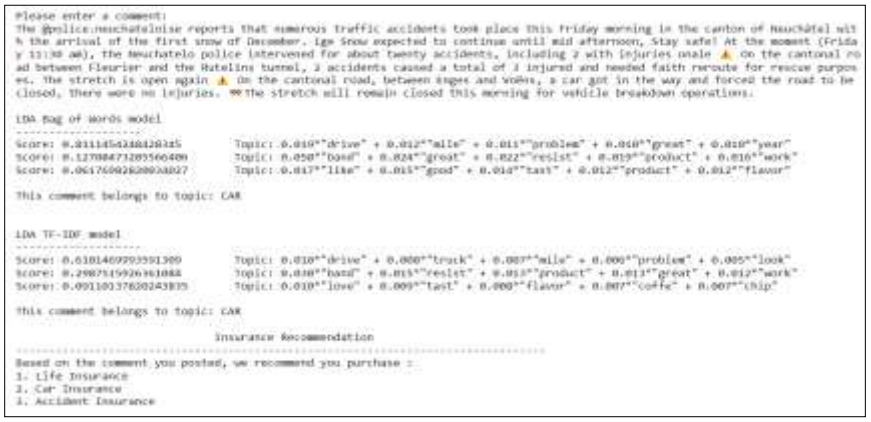

Figure 4. Test case on car accident

Figure 5. Test case on sport comment

As depicted in Table 1, the performance evaluation reveals that the topics of food and cars achieved a perfect score of 5 passes out of 5, while sports attained a slightly lower score of 4 passes out of 5. This discrepancy becomes apparent where the LDA TF-IDF models correctly categorize the comments under the sports topic, however, the LDA bag of words models misclassify the same remarks, placing them under the car topic. The expected alignment of comments with the sports theme is compromised. This discrepancy arises because the TF-IDF model incorporates information on both more and less significant words, providing a nuanced representation, whereas the Bag of Words model merely offers a set of vectors indicating word occurrences in the document (reviews). This fundamental difference in their mechanisms accounts for the divergent results observed in the categorization process.

Table 1. Summary of the test plan

Topic	Results (Number of Passes/Total Test Input)
Food	5/5
Car	5/5
Sports	4/5

Conclusion

In conclusion, this research endeavors to harness the power of social media and machine learning for the efficient identification of potential insurance customers. The study revolves around the development of a comprehensive system that integrates a social media web crawler with advanced algorithms, including Web Content Mining, Noisy Text Filtering, Named Entity Extraction, Part-Of-Speech (POS) Tagging, and Text Clustering. The primary goal is to streamline the extraction of structured information from the vast and heterogeneous pool of unstructured data present on social media platforms. The results showcase the successful implementation of the proposed methodology. Figures 3 to 5 illustrate the system's adeptness in categorizing diverse comments related to food, car accidents, and sports, subsequently recommending relevant insurance types based on the identified themes. The performance evaluation, as depicted in Table 1, reveals commendable scores for the food and cars categories, achieving a perfect score of 5 passes out of 5. However, the sports category attained a slightly lower score of 4 passes out of 5, emphasizing the nuanced challenges in categorizing comments accurately.

An insightful revelation arises from the discrepancy between LDA TF-IDF and LDA Bag of Words models in classifying sports-related comments. The former demonstrates accuracy in identifying sports-related content, while the latter misclassifies these remarks under the car topic. This divergence highlights the impact of the models' mechanisms, where the TF-IDF model's nuanced representation of more and less significant words proves advantageous compared to the Bag of Words model's simplistic approach based on word occurrences.

This research not only contributes to the field of information extraction from social media but also presents a novel approach for insurance companies to identify potential customers

efficiently. The integration of social media data analysis with machine learning provides a valuable tool for insurance companies to navigate the dynamic landscape of customer preferences in the digital age.

Acknowledgements

This research is supported by Tunku Abdul Rahman University of Management and Technology, Malaysia.

References

Akhmadeeva, I., Zagorulko, Y., & Mouromtsev, D. (2016). Ontology-based information extraction for populating the intelligent scientific internet resources. In D. Vrandecic, K. Bontcheva, & R. Klinger (Eds.), *Proceedings of the International Conference on Knowledge Engineering and the Semantic Web (KESW 2016)* (Vol. 649, pp. 119–128). Springer. https://doi.org/10.1007/978-3-319-45880-9_10

Badieh Habib Morgan, M. (2014, August). Information extraction for social media. In *Proceedings of the Third Workshop on Semantic Web and Information Extraction (SWAIE 2014)*. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-6202

Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., & Aswani, N. (2013). TwitIE: An open-source information extraction pipeline for microblog text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)* (pp. 83–90).

Carlson, A., Betteridge, J., Wang, R., Hruschka, E., & Tom, M. (2010). Coupled semi-supervised learning for information extraction. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM 2010)* (pp. 101–110). https://doi.org/10.1145/1718487.1718501

Dong, H., Wang, W., Coenen, F., & Huang, K. (2020). Knowledge base enrichment by relation learning from social tagging data. *Information Sciences, 526*, 339–355. https://doi.org/10.1016/j.ins.2020.04.002

Hu, D., & Xia, Q. (2021). Internet false news information feature extraction and screening based on 5G Internet of Things combined with passive RFID. *Computational Intelligence and Neuroscience, 2021*, Article 9696472, 1–11. https://doi.org/10.1155/2021/9696472

Jones, K., & Sah, S. (2023). The implementation of machine learning in the insurance industry with big data analytics. *International Journal of Data Informatics and Intelligent Computing*, *2*(2), 21–38. https://doi.org/10.59461/ijdiic.v2i2.47

Kim, H., Lee, W., Lee, E., & Kim, S. (2023). Review of evaluation and interpretation method for LDA model. *The Korean Data Analysis Society, 25*(4), 1299–1310. https://doi.org/10.37727/jkdas.2023.25.4.1299

Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights, 1*(2), 100012. https://doi.org/10.1016/j.jjimei.2021.100012

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*, Article 160. https://doi.org/10.1007/s42979-021-00592-x