How can social media inform marketing strategies for life insurance sales

Guan Yue Wang[1]

[1] Johns Hopkins University, 11100 Johns Hopkins Road Laurel, MD 20723, USA
gwang39@jhu.edu

Abstract. understand the needs of the customer and the develop effective marketing has always been one of the most important topics being studied in the life insurance industry. Traditional approaches such as survey and interviews usually come with heavy capital expense and time commitment. Nevertheless, the data can become outdated over time with the shift in social preference as well as technology advancement. Social media provides a rich source of comments and discussions from both user and marketer perspectives. This paper uses sentiment analysis and n-gram language model to develop an information retrieval approach in order to extract meaningful words and phrases from the reddit data in the attempt to drive actionable marketing strategies. With the methods established, algorithms can be ran on a continuing basis to capture any movements in the industry in terms of customer needs and emerging marketing techonologies.

Keywords: life insurance, marketing, social media, reddit, natural language processing, sentiment analysis, n-gram language model.

1 Introduction

Developing a keyword(s) retriever for user comments with sentiment considerations can be a valuable technique for companies to establish effective marketing campaigns and improve sales. This can be especially important for the life insurance industry due to the nature of the product. To be specific, there are mainly two types of life insurance products – whole life and term life, one that people only buy once in their life time whereas the other only require one purchase with renewal every 10 to 20 years assuming they don't switch companies. As a result, discussions for life insurance products can be hard to find in contrast to commercial goods where people constantly replace old devices and share feedbacks everywhere. Some potential market research solutions are conducting survey and interviews but they are usually expensive in terms of capital expense and time commitment. Nevertheless, the data obtained can be outdated over time and are exposure to different biases such as observer-expectancy effect, confirmation bias [1], and user privacy concerns. Meanwhile, there are massive amount of life insurance and marketing related discussions available on social media platform which can potentially provide great insights into customer needs, social preference, product feedbacks, and marketing techniques. However, key words and

meaningful phrases cannot be easily identified without a rigorous natural language processing algorithm, this paper investigates methods to retrieve valuable keywords and meaningful short phrases towards research topics by mining and analyzing the reddit user comments.

Reddit is a social media platform and online community that allow members to share contents such as text posts, links, images, and opinions towards different topic topics called subreddit. As of 2018, there are 1.2 million of subreddits available covering a variety of topics including news, science, movies, music, food, games, fitness, etc. [2]. According to Alexa Internet, Reddit had 542 million monthly visitors (234 million unique users), ranking as the No. 6 most visited website in U.S. and No. 21 in the world As of March 2019, with 53.9% of its user base coming from the United States, 8.2% from UK, and 6.3% from Canada[3].

With such strong user basses and rich source of comments, thousands of discussions, onions, and ideas, on different topics are shared on reddit every day. Consequently, there are tremendous amount of opportunities filtering relevant information and turning that into actionable insights so that companies can learn from it and use it to drive more effective marketing strategies and improve sales.

2 Background

This paper focuses on building a keyword(s) retriever with sentiment considerations to drive effective digital marketing in the life insurance industry. An effective marketing can be established from two dimensions, one is marketing contents, the other is marketing techniques. This paper extract both types of the information from reddit by filtering discussions among various subreddits.

The data used in this paper is extracted from posts and their comments within various subreddit community. To be more specific, digitalmarketing subreddit contributes directly to our digital marketing comments research while insurance related comments are identified and filtered from various finance and insurance related subreddits including insurance, personalfinance, and personalfiancecanada. Within digitalmarketing subreddit, a lot of discussions are going on everyday discussing the pros and cons of marketing strategies, evaluating digital marketing tools, sharing experience on different marketing platforms. On the other hand, people frequently seek for help regarding life insurance related questions. Within the discussion, the drivers behind life insurance purchase as well as concerns related to the purchasing decisions are usually shared and discussed. As a result, keywords from these comments would improve company's understanding of the topics.

While extracting keywords from the text is not a tough task, it is hard to identify the relationship between the keywords and the research topics. Therefore, a sentiment classifier is layered in to categorizes the comments prior to our keyword(s) extractions. For our purpose, NLTK's naïve Bayes classifier is leveraged to help classify our comments into positive and negative sentiment groups [4]. Consequently, we

were able to distinguish between positive and negative comments before extracting the keywords and interpret it regarding our research topics.

Training data for sentiment analysis is obtained from Sentiment140 which is developed to discover sentiment of brands, topics, or products on twitter. This is selected as there is no sentiment training data available specific to reddit while twitter is a similar social media platform as reddit. The Sentiment140 corpus is particularly useful for use cases such as brand management and purchase planning which aligns with our research interests on Reddit [5]. Quality of the sentiment analysis is measure through accuracy score which can be calculated as:

Accuracy = Number of Correct Predictions / Total Number of Predictions Made (1)

An ideal accuracy would be 70% which is considered doing nearly as well as humans for sentiment classifications [6].

3 Method

Research consists of three phases. The phase one starts with finding relevant information on reddit. The second phase involves splitting the collected comments into positive and negative sentiment groups. The third phase focuses conducting n-gram model on two groups to identify key words or phrases as positive and negative factors towards our research topics. Finally, top x of the positive and negative n-gram words/phrases are populated for further analysis and interpretations.

During the first phase, an access token has to be granted as the set up for the Reddit data API. Next, the relevant subreddits have to be identified, it can to be a subreddit directly links to our topic or a subreddit contains a lot of relevant information. After that, python script is used to extract the top 1000 posts and replies within the posts as csv for further analysis. Note customized keyword filter within the post titles might be required if we are not working with a directly related subreddit community. For example, ‘life insurance’ has to be spotted in the post title in order for the post to be included when we extract data from personalfinance subreddit.

After the raw data is ready, sentiment analysis is conducted on the data to split the comments into positive sentiment group and negative sentiment group. NLTK’s sentiment analysis package is leveraged to identify the sentiment of the comments. To be more specific, the algorithm uses Liu and Hu opinion lexicon which simply counts the number of positive, negative and neutral words in the sentence and classifies it depending on which polarity is more represented. (Words that do not appear in the lexicon are considered as neutral.) [7] With the sentiment classifier splitting the comments into positive and negative sentiment groups, we can better understand if a word or phrase is positively or negatively related to our research interests.

Next, the n-gram model is applied to our positive and negative sentiment data. n-gram is a contiguous sequence of n items from a given sample of text and it is usually used to model natural language sequences using the statistical properties of n-gram

[8]. For our purpose, we use it to identify the most common unigram as well as the frequent words around it. The n-gram model improves the retrieval performance for our keywords identification process as well as provides insights into the probability keywords appear in out interested topic. Combined with sentiment tag, we can then understand whether the specific keywords/phrases are positively or negatively related to our topics.

In addition to understand the pure positive and negative sentiment of the keywords, domain knowledge is required to validate the results and further interpret the outcomes. Consequently, insights and knowledge come out of the algorithm are expected to be turned into actionable sales and marketing strategies.

Findings

3.1 Digital Marketing Related Subreddit Comments

After 19122 comments analyzed through the algorithm, 78% are categorized as positive sentiment while the other 22% are tagged as negative sentiment. This demonstrates most of the ideas, comments, and feedbacks shared in the digital marketing subreddit can be more related to positive experience.

The results of the keyword(s) retriever can be found in Table 1 (positive sentiment) and Table 2 (negative sentiment). Let's start by looking into some common words in both tables. We see there are a lot of google related terms illustrating the popularity of the marketing platform as well as its mixed feedbacks. Also, a lot of marketing methods exist in both tables such as email marketing, content marketing, digital marketer, google ads, search engine, search engine optimizations. By obtaining this list, company can learn what marketing strategies are currently being deployed in the market and evaluate each of them respectively. Next, let's investigate some terms only exist in one sentiment group. To start with, we see Facebook and Facebook ads only appear in the positive sentiment table, this demonstrates feedbacks related to Facebook platform tend to be more positive and can potentially represent that it is a better platform compared to google which has mixed feedbacks. In addition, we see terms such as learn and digital marketing courses only show in the positive sentiment which demonstrates people are more motivated to learn in digital marketing and a lot of these learning experience comes with positive comments.

With the results in Table 1 and 2, company would be able to know what marketing tools and approaches are on the trend as well as which of them comes with good user feedbacks. In addition, any change in the popularity of the tool and the shift in sentiment can be captured by running the algorithm on a continuous basis. As a result, life insurance company would always have access to a list of popular marketing tools/methods so that they can evaluate and deploy the most effective marketing campaign as well as turn that into a positive impact on their sales.

Table 1. Most Common Positive Sentiment N-gram In Digital Marketing Related Comments

Unigram	Count	Bigram	Count	Trigram	Count
marketing	1421	digital marketing	560	social media marketing	560
digital	813	social media	299	digital marketing course	299
google	751	google ads	108	part digital marketing	108
content	572	google analytics	100	digital marketing strategy	100
ads	516	email marketing	91	marketing digital marketing	91
social	490	content marketing	79	seo social media	79
facebook	478	facebook ads	59	message compose r	59
seo	471	digital marketer	56	google tag manager	56
time	381	landing page	54	digital marketing manager	54
learn	359	search engine	47	social media examiner	47

Table 2. Most Common Negative Sentiment N-gram In Digital Marketing Related Comments

Unigram	Count	Bigram	Count	Trigram	Count
marketing	1160	digital marketing	507	social media marketing	560
digital	702	social media	259	marketing digital marketing	299
google	399	google analytics	65	digital marketing manager	108
social	380	google ads	57	marketing social media	100
time	364	digital marketer	49	search engine optimization	91
media	323	email marketing	37	learn digital marketing	79
ads	304	content marketing	36	digital marketing agency	59
help	246	search engine	32	social media platform	56
way	239	media marketing	31	job digital marketing	54
seo	238	landing page	29	digital marketing digital	47

3.2 Life Insurance Related Subreddit Comments

After 3035 life insurance related comments analyzed through the algorithm, 55% are categorized as positive sentiment whereas the other 45% are tagged as negative sentiment. The fairly even split gives us sufficient data to investigate both positive and negative drivers behind life insurance purchasing decisions.

The results of the keyword(s) retriever can be found in Table 3 (positive sentiment) and Table 4 (negative sentiment). Let's start by looking into some common words in both tables. Unsurprisingly, there most common words and phrases are different insurance policies and insurance companies, this is a result of people discussing and seeking feedbacks on various types of life insurances as well as companies. One of the highlighting factors driving the life insurance purchases are family status based on the frequent appearance of related words such as family, wife, and family members.

In addition, cash value is also a phrase comes up in both tables which can be another important consideration for potential buyers. Next let's look at some phrases unique to each table. In the positive sentiment table, we see several instances of financial planner and financial advisers which might suggest these financial service representatives brings people positive experience and can potentially help drive the purchasing decisions. On the other hand, there are several time related terms such as years, 20 year term, 30 year term which can mean the length of the benefit could be one of the concerns stopping people from buying life insurance.

With the results in table 3 and 4, life insurance companies can incorporate some of the key factors such as family consideration and cash value into their marketing contents for a more convincing marketing campaign. Meanwhile, assistance from financial planner and advisors can also be leveraged as it seems to bring positive experience to the clients.

Table 3. Most Common Positive Sentiment N-gram In Life Insurance Related Comments

Unigram	Count	Bigram	Count	Trigram	Count
insurance	169	life insurance	84	life insurance policy	11
life	154	whole life	34	whole life policy	8
policy	117	social security	15	beneficiary life insurance	6
money	88	insurance policy	14	1 2 million	5
pay	78	insurance company	12	whole life policies	4
trust	59	credit card	12	amount received financial	4
help	51	life policy	11	received financial adviser	4
financial	51	cash value	10	financial adviser help	4
wife	50	financial planner	9	adviser help percentage	4
family	44	family members	8	life insurance payout	3

Table 4. Most Common Negative Sentiment N-gram In Life Insurance Related Comments

Unigram	Count	Bigram	Count	Trigram	Count
life	485	life insurance	233	life insurance policy	35
insurance	439	whole life	118	whole life policy	35
policy	373	life policy	59	whole life insurance	12
money	219	insurance policy	45	term life insurance	11
term	178	death benefit	42	term life policy	10
years	159	term life	41	life insurance company	9
whole	138	cash value	36	life insurance policies	9
pay	134	term policy	33	20 year term	7
family	120	insurance company	29	life insurance agent	5
trust	120	social security	23	20 30 years	5

4 Conclusion

Implementing a Reddit keyword retriever utilizing sentiment analysis and n-gram language model appears to provide some great insights into the insurance marketing strategies for sales improvement. It helps identify some of the popular tools in the field of digital marketing and sentiments of the feedbacks behind them. On the other hand, it reveals some of the key factors driving the life insurance purchasing decisions such as family status, length of the life insurance policy, price of the benefit, and types of insurances. Furthermore, sentiment analysis further explains the positive/negative relationship between the drivers and the life insurance purchasing decisions.

There are various limitations and challenges in this data mining and keyword retrieval process. First is the data availability, this can be addressed in terms of both API data volume restrictions and nature of the data. To start with, Reddit API imposed a 1000 posts request limit so that we can only obtain top 1000 posts and their comments for one subreddits. Moreover, life insurance related posts require customized filtering on the comments which further reduce the data volume. In addition, the data can only be considered a fair representation of the Americans due to the fact that more than 50% of the reddit users are from United States[10]. Secondly, the sentiment analysis is conducted with various limitations and assumptions. As we use Sentiment 140 corpus for sentiment analysis, we assume reddit data has the same attributes as twitter data. This is not always true because the length of the comments and user base are different in two social media platforms. Also, things such as sarcasm, hyperbole, and punctuations cannot be fully interpreted by the sentiment classifier. Moreover, only 10000 of the sentiment 140 corpus are used in the training data due to computing power constrains which further limits accuracy of the sentiment classification. Last but not the least, reddit specific stop words are manually identified so our results can potentially be impacted this.

Fortunately, some of these limitations can be solved so that the method can be further improved in future. For example, the volume of the data can be increased by utilizing or more advanced data API so that we can incorporate more comments into our algorithm. With more data, we might be able to develop our own training corpus to the sentiment classifier as well as stop word list so everything is more customized for Reddit data. Consequently, more valuable information can be extracted from the reddit comments which can lead to more effective marketing campaigns for insurance companies.

References

1. Goldstein, Bruce. "Cognitive Psychology". pp. 374 Cengage Learning (2011)
2. Statista Infographics, The Explosive Growth of Reddit's Community, https://www.statista.com/chart/11882/number-of-subreddits-on-reddit/, last accessed 2019/05/12.
3. Alexa Internet, Reddit.com Site Info, https://www.alexa.com/siteinfo/reddit.com, last accessed 2019/05/12.
4. NLTK, Learning to Classify Text, https://www.nltk.org/book/ch06.html, last accessed 2019/05/12.
5. Sentiment140, A Twitter Sentiment Analysis Tool, http://help.sentiment140.com/home, last accessed 2019/05/12.
6. Roebuck, K. Sentiment Analysis: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Lightning Source (2012)
7. nltk.sentiment package, https://www.nltk.org/api/nltk.sentiment.html, last accessed 2019/05/12.
8. Andrei Z., Steven C., Mark S., Zweig, Geoffrey: Computer networks and ISDN systems, vol. 29, iss. 8-13, pp. 1157-1166 . Elsevier Science (1997)
9. The Demographics of Reddit, https://www.techjunkie.com/demographics-reddit/, last accessed 2019/05/12.