BRAND REPUTATION MONITORING SYSTEM BASED ON SENTIMENT ANALYSIS USING THE K-NEAREST NEIGHBOR METHOD

Bambang Wisnu Widagdo[1], Mochamad Adhari Adiguna[2], Anggraeni[3]

[1,2,3]Department of Informatics, Universitas Pamulang, Banten, Indonesia, 15310
e-mail: *[2]dosen01864@unpam.ac.id

Abstract

This research paper aims to describe the research and implementation of a brand reputation monitoring system. The system, called Monitool, is needed by the digital marketing division of PT Lingkar9 Titian Media to measure the brand reputation of its clients. Monitool focuses on monitoring and analyzing a brand on social media and news portals. With Monitool, the digital marketing division of PT Lingkar9 Titian Media can easily make digital marketing plans and can evaluate brand reputation efficiently. Currently the marketing division of PT Lingkar9 Titian Media measures brand reputation using paid tools to measure brand reputation, but it is quite costly, because the costs are not proportional to the profits, especially if the project being handled suffers a loss. Monitool uses the K-Nearest Neighbor (KNN) method to measure the reputation of a brand. Before measuring reputation, monitool will look for news or social media posts that offend the client's brand. After the data is collected, the data will be processed using the KNN method. After being processed, the system will issue a resume of sentiments consisting of negative, positive and neutral sentiments. The accuracy result of the current test data prediction is 81%. This prediction result will change when the training data is increased or decreased.

Abstrak

Penelitian karya tulis ini bertujuan untuk memaparkan penelitian dan implementasi sistem monitoring reputasi merek. Sistem yang bernama Monitool ini dibutuhkan oleh divisi digital marketing PT Lingkar9 Titian Media untuk mengukur reputasi merek kliennya. Monitool berfokus pada monitoring dan analisis suatu merek di media sosial dan portal berita. Dengan Monitool, divisi digital marketing PT Lingkar9 Titian Media dapat dengan mudah membuat perencanaan marketing digital dan dapat mengevalusi reputasi merek secara efisien.

Saat ini divisi marketing PT Lingkar9 Titian Media mengukur reputasi merek menggunakan perangkat berbayar untuk mengukur reputasi merek, tetapi cukup menguras biaya, karena biaya nya yang tidak berbanding dengan keuntungan yang didapat, apalagi jika proyek yang ditangani merugi.

Monitool menggunakan metode K-Nearest Neighbor (KNN) untuk mengukur reputasi suatu merek. Sebelum mengukur reputasi, monitool akan mencari berita atau postingan sosial media yang menyinggung merek klien, setelah data terkumpul data akan diolah menggunakan metode KNN, setelah diolah sistem akan mengeluarkan resume sentimen-sentimen yang terdiri dari sentimen negatif, positif dan netral. Hasil akurasi dari prediksi data uji saat ini adalah sebesar 81%. Hasil prediksi ini akan berubah ketika data latih bertambah atau berkurang.

Keywords: *Brand reputation monitoring, Brand reputation measuring application, Sentiment analysis, K-Nearest Neighbor*

1. INTRODUCTION

The development of technology is now increasing rapidly and many new innovations have emerged. There are so many tools available in cyberspace, from the fields of entertainment, economics, education, social and much more. All offline information has been transformed into digital information. Therefore, the use and

processing of data has also changed. The use of media is very important for business people. After marketing a product from a brand, it is important for the owner to evaluate the marketing results. The marketing results will later be used as a reference to improve future marketing. Likewise with digital marketing vendors, they need tools to measure their clients' brand reputations that are efficient and effective. Currently, there are many brand monitoring tools scattered in cyberspace. The prices are varied and competitive. There is also a free trial, but the available features are limited.

PT Lingkar9 Titian Media is one of the companies engaged in the creative industry technology. One of its divisions is the digital marketing division which also requires tools to measure brand reputation. Currently, the digital marketing division of PT Lingkar9 Titain Media uses paid reputation measuring tools. However, after being evaluated, the use of these tools is quite costly, especially if the project being carried out suffers a loss. Therefore, PT Lingkar9 Titian Media took the initiative to create the Monitool project. Monitool is a client brand reputation monitoring system that will be owned exclusively by PT Lingkar9 Titian Media. Monitool can search for news containing the brand name mentioned on social media and portals. In the background, monitool will process the search data results into sentiments that are divided into negative, positive and neutral sentiments, using the K-Nearest Neighbor (K-NN) algorithm method.

Here we are given the trust by PT Lingkar9 Sembilan Titian Media to continue the development of the Monitool project concept of brand reputation monitoring system based on sentiment analysis using the K-Nearest Neighbor method.

2. LITERATURE REVIEW

2.1. Sentiment Analysis

Sentiment analysis is the process of understanding, extracting and processing text data automatically to obtain labels or sentiment information contained in an opinion sentence. This analysis is carried out to see the tendency of public opinion to a problem, object or a brand, whether it tends to be positive, negative or between the two (neutral) [1].

2.2. K-Nearest Neighbor

K-nearest neighbor (KNN) is the simplest classification method to apply the data classification method, compared to other classification methods. The kNN method is used to perform predictive analysis of unknown data or it is difficult to determine the probability [2].

Leif E. Peterson suggested that the kNN method is a supervised classification method, which uses labeling on training data. As for the unsupervised classification, the cluster method or similar methods do not label the training data [2].

Determination of this kNN prediction uses K, where K is the parameter of the number of neighbors that will be used for predictive analysis. According to Leif E. Peterson, the kNN classification formula commonly used is Euclidean distance [2]. Here's the formula for the euclidean distance:

$$d(x_a, x_b) = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2 + \ldots + (x_{an} - x_{bn})^2} \quad (1)$$

2.3. Text Preprocessing

Text Preprocessing is data processing in the form of text which will be converted into numbers which can later be used to determine the prediction of data classification with text. This text preprocessing decision aims to make text prediction analysis simpler and this text preprocessing also does not affect the prediction results in the next model [3]. According to the policy of the results of the analysis of political texts, there are word sequences that need to be removed, then leaving a collection of basic words, in order to optimize the text prediction analysis process [4].

2.4. Supervised Classification Method

The supervised classification method is a classification method that has directions. This method is based on the existence of a dataset which sample data has existed since the beginning and has a class label [5]. The advantages of this supervised method are that it has control over the data class based on the training sample, and the control over the accuracy of the classification. The drawbacks are forced data interpretation, the selection of training samples is not necessarily representative, and the presence of unidentified spectral classes [6].

2.5. Scraping Method

The scraping method is one way to separate the main content of a web page from parts that are not related to the content. The content taken can be in the form of the title of the site page, the date listed on the site's article page or whatever is on the site can be taken in the form of parts [7].

2.6. Unified Modeling Language Diagrams

The Unified Modeling Language (UML) is a standard diagram format for designing and analyzing software projects. The most important format in the UML component is the class diagram, which models information in the domain related to objects that have been arranged both classes and relations [8].

2.7. Database

The database can be thought of as a filing cabinet. Base is defined as a headquarters or warehouse, nesting headquarter or gathering place [9]. Data is a value that also represents a description of an object or event. A well-designed database, looping or redundant data can be eliminated so that inconsistent data storage can be minimized [9].

3. METHOD

The following are the research methods and framework of "Brand Reputation Monitoring System Based On Sentimen Analysis" which we describe in chart form:

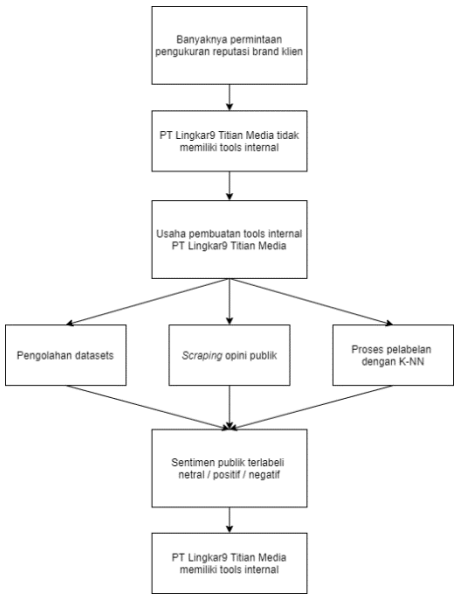

Figure 1. Framework

This image describes the decision-making process for this brand reputation measurement application. Starting from the many requests for brand reputation measurement from clients of Lingkar9's digital marketing division, and at that time Lingkar9 still relied on external tools and of course the cost of using external tools was very expensive. Finally, Lingkar9 decided to make an internal application, and the author was given the trust to build this application together. The author conducted research, and the results describe several text processing methods, namely: dataset processing, public opinion scraping and finally the labeling process (positive/negative/neutral) opinion with KKN.

4. RESULTS AND DISCUSSION

4.1. Analysis Of Current System Sistem

The current system analysis is similar applications that PT Lingkar9 Titian Media wants to make, such as brand24, mediatoolkit, socialbakers and many other digital-based monitoring media. What we are currently doing is brand24 and mediatoolkit, and from the results of the research, we make the flow of similar programs from both is as illustrated in the following figure:

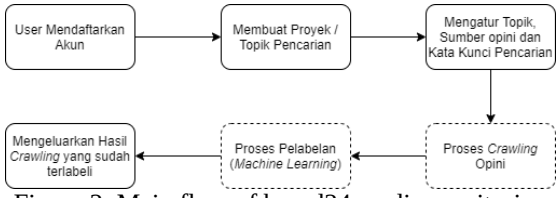

Figure 2. Main flow of brand24 media monitoring program and mediatoolkit

This image describes the main flow of this brand reputation measurement application. Starting from users/clients who register their accounts in the application until users get labeling results from their brand reputation.

It can be seen in Figure 1 that there are differences between the two, but not much. And another similarity is setting up social media accounts to retrieve analytic information from Facebook and Instagram, by authorizing user accounts first. There is a difference with the media monitoring system created by PT Lingkar9 Titian Media. The main concept or flow is the same, namely setting search topics, crawling data from various sources, labeling data and issuing labeling results. For this first phase, we build an application by going through the main flow of this media monitoring system, but still being able to crawl data from twitter and instagram by scraping method using the python and PHP programming languages. Then the labeling process is carried out using the supervised learning classification method, here we use K-nearest neighbor.

4.2. Data Analysis

The data analysis is done by understanding the concept of the system, then searching for datasets in the form of reviews or opinions. Finally we found several kinds of datasets, and we took stable datasets more detailed description of the review from Claudhiah's github account, and processed independently by Chlaudiah Julinar and

Vina Fadriani Effendi. These datasets are obtained from the crawling process on the femailydaily.com site. The number of datasets reaches 500 Indonesian language data, which we think is sufficient for training data during the machine learning process.

4.3. Entity Realtionship Diagram

Entity Realtionship Diagram or ERD is used for modeling data structures and relationships between data [10]. This monitoring system has several entities, namely:

1. Client: client id, company name, company email, company phone number
2. Project: project id, project name, keywords, opinion sources
3. Crawling: crawling id, project id, crawl date, keywords, accuracy (%), best K
4. Datasets (Bank Sentiment): sources of opinions, opinions, labels
5. Internal User (Administrator): user id, name, email, phone number, access level.
6. Opinion labeling: id_opinion, project id, crawl id, opinion, source, publish date, label, conversion status

And next is the relationship between entities which is described as follows:

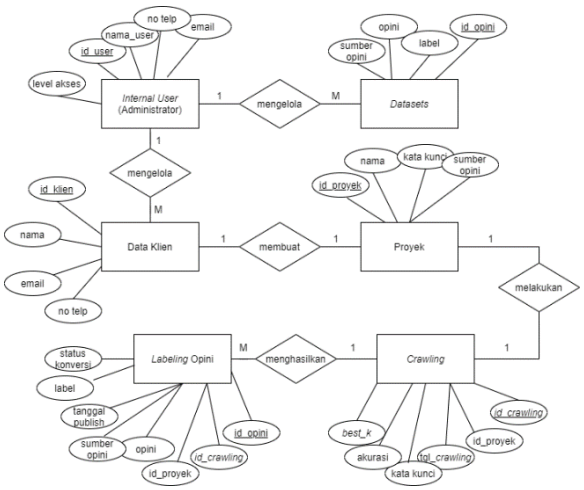

Figure 3. ERD brand reputation monitoring system

This figure describes the relationships between the table entities present in the application. From internal users (administrators) who manage data sets and account data of client. Afterward, the client creates the project (information about the brand), then the system crawls the project data, the crawling results will proceed to the labeling process.

4.4. Use Case Diagram

Use case diagram is depicted in Figure 4 below.

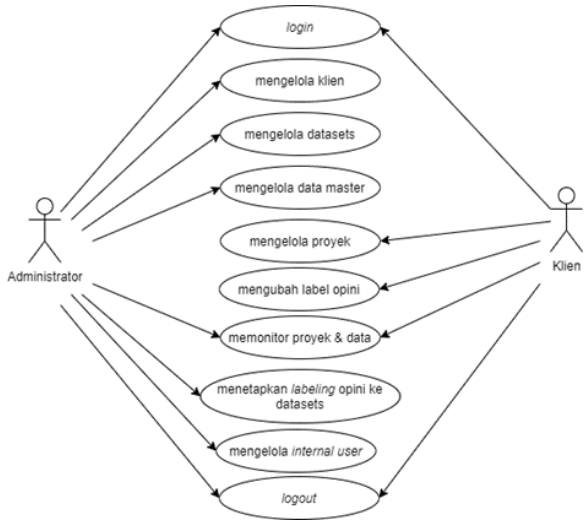

Figure 4. Use case diagram of brand reputation monitoring system

This use case describes the interaction between 2 levels of users in this brand reputation measurement application.

4.5. Sequence Diagram

Sequence diagram is demonstrated in Figure 5 - 7.

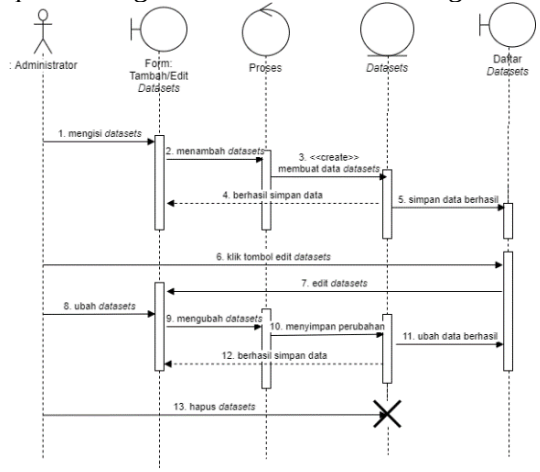

Figure 5. Sequence diagram of datasets/sentiment bank

Describes the process flow for adding, modifying, and deleting a sentiment data set.

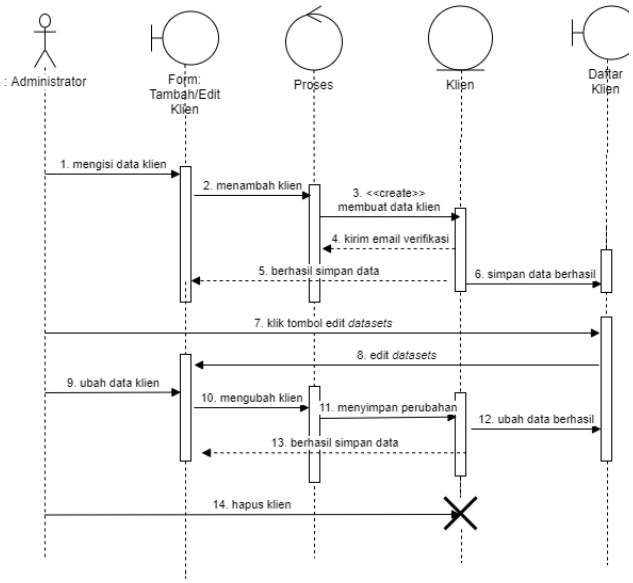

Figure 6. Sequence diagram of client management

Describe the process flow of adding, changing, and deleting client accounts.

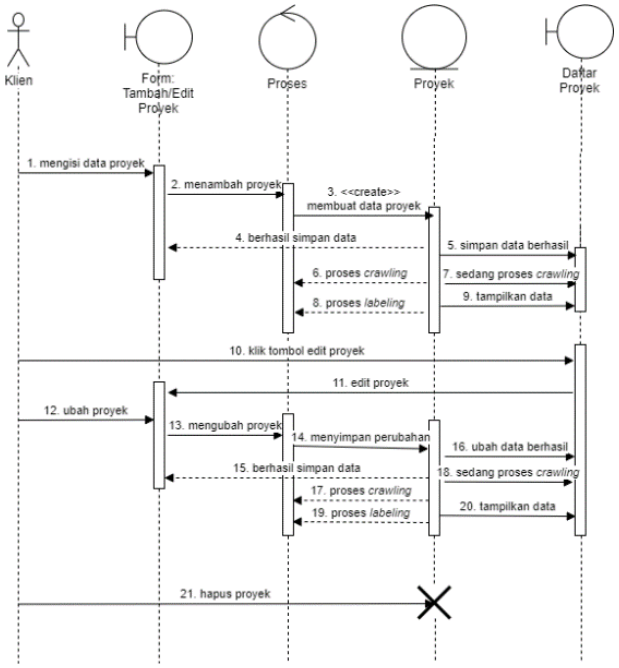

Figure 7. Sequence diagram of client management

Describe the process flow of adding, changing, and deleting client project data.

4.6. Activity Diagram

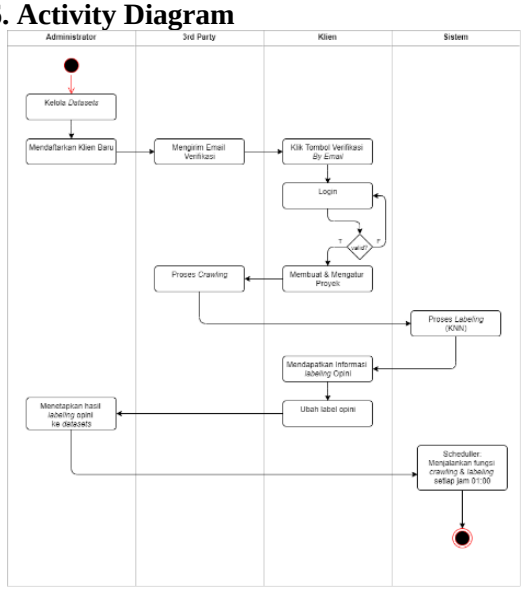

Figure 8. Activity diagram of brand reputation monitoring system

Describes the main process flows in this brand reputation measurement application.

4.7. User Interface

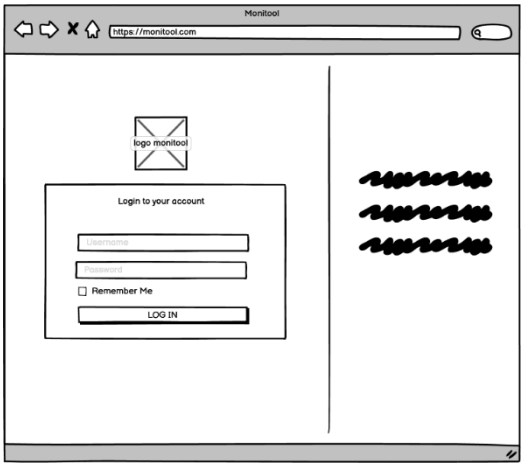

Figure 9. UI Mockup – login page

This image is a UI Mockup for the login page.

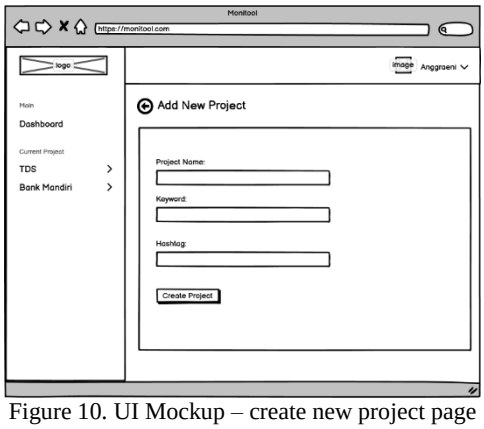

Figure 10. UI Mockup – create new project page

This image is a UI mockup for the form of creating new project, which present on the client dashboard.

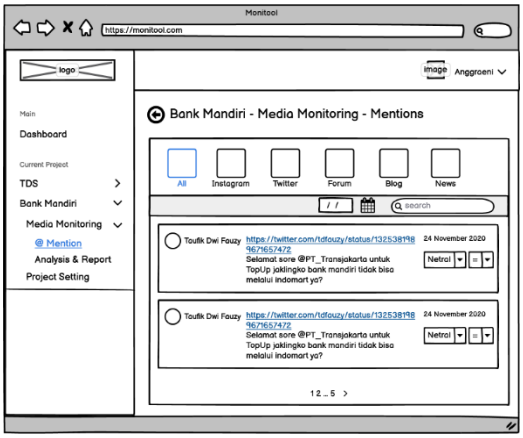

Figure 11. UI Mockup –detail page of project after crawling

This image is a UI mockup for a page displaying project data or brand, which has been crawled and labeled.

Monitool uses the K-Nearest Neighbor (KNN) method to measure the reputation of a brand. Before measuring reputation, monitool will look for news or social media posts that offend the client's brand. After the data is collected, the data will be processed using the KNN method. After processing, the system will issue a resume of sentiments consisting of negative, positive and neutral sentiments. The accuracy result of the current test data prediction is 81%. This prediction result will change when the training data is increased or decreased.

5. CONCLUSION

Based on Sentiment Analysis With the K-Nearest Neighbor Method, it can be concluded that the digital marketing division of PT Lingkar9 Titian Media can have a special brand reputation monitoring system, without using paid tools in cyberspace. In addition, with the machine learning method, K-NN, opinion labeling will be carried out automatically by machine learning which previously learned from available datasets or training data.

6. ACKNOWLEDGEMENTS

This research was supported by the Department of Informatics of Pamulang University. We thank our colleagues from the Department of Informatics of Pamulang University who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

7. REFERENCES

[1] Imam Fahrur Rozi, Sholeh Hadi Pramono, Erfan Achmad Dahlan, “Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi”, Jurnal EECCIS Vol. 6 No. 1, 37-43, 2012.

[2] Peterson, L. E., “K-nearest neighbor. Scholarpedia, 4(2), 1883”, 2009.

[3] Denny, M., & Spirling, A., “Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. When It Misleads, and What to Do about It (September 27, 2017)”., 2017

[4] Grimmer, J., & Stewart, B. M., “Text as data: The promise and pitfalls of automatic content analysis methods for political texts”., Political analysis, 21(3), 267-297, 2013.

[5] Agusta, Yudi, 2008, TIPS: Pembuatan Literature Review, http://yudiagusta.wordpress.com/2008/04/08/tips-pembuatan-literature-review

[6] Rosi Septiani, I Putu Ananda Citra, A Sediyo Adi Nugraha, “Perbandingan Metode Supervised Classification dan Unsupervised Classification terhadap Penutup Lahan di Kabupaten Buleleng”, Jurnal Geografi 16(2), 90-96. 2019.

[7] Mardi Siswo Utomo, Setyawan W., PERANCANGAN APLIKASI WEB SCRAPING UNTUK KOLEKSI KONTEN RESEP MASAKAN TRADISIONAL JAWA BERBASIS XML,2015. https://garuda.ristekbrin.go.id/documents/detail/387106

[8] Berardi, D., Calvanese, D., & De Giacomo, G., “Reasoning on UML class diagrams. Artificial intelligence”, 168(1-2), 70-118., 2005

[9] Kurniawati, Rita, “Sistem Basis Data”, Unikom, 2007.

[10] Eriana, E. S., “Pengujian Sistem Informasi Aplikasi Perpustakaan Berbasis Web Dengan White Box Testing”. In ESIT (Vol. 15, No. 2, pp. 28-33), 202