EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI: 10.20885/EKSAKTA.vol2.iss1.art3

Research Article

Hidden Markov Model for Sentiment Analysis using Viterbi Algorithm

Nursyiva Irsalinda[1*], Haswat[2], Sugiyarto[3], Meita Fitrianawati[4]

[1,2,3] Mathematics Department, Universitas Ahmad Dahlan, Jalan Kolektor Ring Road Selatan, Tamanan Banguntapan Bantul Yogyakarta
[4] Department of Elementary School Education, Faculty of Teacher Training and Education, Universitas Ahmad Dahlan
[*] *Corresponding author: nursyiva.irslinda@math.uad.ac.id*

Received: 25 November 2020; Accepted: 7 January 2021; Published: 12 January 2021

Abstract: Data mining is an activity to extract the knowledge from large amounts of data as a very important information. The type of data in the era of 4.0 is data in the form of text, which is very much derived from social media. Recently, text becomes very important in some applications, such as the processing and the conclusion of a person's review and analysis of political opinion which is very sensitive in almost all countries, including Indonesia. Online text data that circulating on social media has several shortcomings that could potentially hinder the analysis process. One of the drawbacks is the people can post their own content freely, so the quality of their opinions cannot be guaranteed such as spam and irrelevant opinions. The other drawback is the basic truth of the online text data is not always available. Basic truth is more like a particular opinion, indicating whether the opinion is positive, negative, or neutral. Therefore, the main objective of this study is to improve the forecasting accuracy of online text data analysis from social media. The method used is Hidden Markov Model (HMM) with Viterbi Algorithm that applied to extract the dataset sentiment at the 2015 elections in Surabaya from the popular site micro blogging called Twitter. The result of the study is Viterbi algorithm has predicted the best route with the candidate Tri Rismaharini gained a prediction of neutral sentiments, whereas Rasiyo candidates gained sentiment negative predictions as well. The proposed Model is accurate to predict candidate features. It also helps political parties to introduce candidates based on reviews so that they can increase candidate performance or they can manage broad publicity to promote candidates.

Keywords: Hidden Markov Model, Stochastic, Sentiment Analysis

Introduction

Data mining is an activity to extract the knowledge from large amounts of data as a very important information. In general, data mining tasks can be classified into two categories: descriptive and predictive. The task of extracting or mining descriptively is to classify the general nature of a data in the database. The predictive Data Mining task is to take conclusions on the last data to make predictions [1].

Hidden Markov Model (HMM) is a statistical model in which a system is being modeled as a Markov process in an unobserved state. On the usual Markov Model, each subsequent state relies on its previous state, this model will show all possible probability between states. Therefore, the probability of transitioning between state becomes the only observed parameter. Markov models are often used for pattern recognition and making predictions. HMM can also be used to find effects on any candidate. Thus, the sequence of steps made by HMM provides an information about the order of the state [2].

The type of data in the era of 4.0 is data in the form of text, which is very much derived from social media. In recent years, natural language processing studies have become more oriented toward opinion mining in social media [3]. Sentiment analysis plays an important role to classify text data into positive, negative, and neutral opinion categories to express opinions in reviews. This process is studied and applied to users who do not explicitly express their sentiment orientation in a particular context [4]. Sentiment analysis has a level of difficulty, among which are assessments expressed in an opinion or part of an opinion addressed to the subject or object, and whether the expressed opinion is positive, negative, and neutral. Recently, text becomes very important in some applications, such as the processing and the conclusion of a

person's review and analysis of political opinion which is very sensitive in almost all countries, including Indonesia. Online text data that circulating on social media has several shortcomings that could potentially hinder the analysis process. One of the drawback is the people can post their own content freely, so the quality of their opinions cannot be guaranteed such as spam and irrelevant opinions. The other drawback is the basic truth of the online text data is not always available. Basic truth is more like a particular opinion, indicating whether the opinion is positive, negative, or neutral [5].

Unstructured data is data that has no specific format or model. Text data, image data and video data are some of the examples of unstructured data. This type of data is estimated to represent 80 percent of the valuable information for most of the organizations [6]. Social media is not only one popular place to talk about a problem, but it is also a place to gather community sentiments about something that is considered viral in the form of text opinions, images or videos [7]. In Twitter the people can post their own content that the quality of their opinions cannot be guaranteed. Therefore, the main objective of this study is to improve the forecasting accuracy of online text data analysis from twitter. The method used is Hidden Markov Model (HMM) with Viterbi Algorithm that applied to extract the text data in Twitter. The Viterbi algorithm proposed by Andrew J. Viterbi in 1967, is a dynamic programming algorithm that finds the most probable sequence of hidden states, called the "Viterbi path" from a given sequence of observed events in the context of a hidden Markov model (HMM) [8]. Viterbi algorithm (VA) on time frequency (TF) distribution is a highly performed instantaneous frequency (IF) estimator [9].

Regional head elections or pemilihan kepala daerah(Pilkada) in a country that adheres to democracy can be held periodically. A political figure who wants to run as a candidate for head of a certain area will see or consider their popularity based on the opinion of the public. The 2015 General Election for Mayor of Surabaya was held on December 9, 2015 to elect the Mayor of Surabaya for the 2016–2021 period. The implementation of this general election coincided with the implementation of simultaneous regional head elections throughout Indonesia on December 9, 2015. There were two pairs of candidates competing in this general election, namely the incumbent pair Tri Rismaharini/Whisnu Sakti Buana which was promoted by the Partai Demokrasi Indonesia Perjuangan (PDI- P) and Rasiyo/Lucy Kurniasari who are promoted by the Partai Demokrat and the Partai Amanat Nasional (PAN). The general election was won by the Tri Rismaharini/Whisnu Sakti Buana pair carried out by the PDI-P with a total vote of 893,087 (86.34%) in accordance with the decision of the Surabaya City KPU on December 22, 2015.

Therefore, this study was conducted to see the sentiment analysis of the two candidates using HMM modeling with the Viterbi algorithm. Thus, it can be concluded whether the results of the sentiment analysis carried out are in accordance with the results of the Surabaya Pilkada in 2015.

Materials and Methods

This study uses Hidden Markov models to foresee the future by considering the hidden problems affected in certain elections 2015 data (reviews on candidates at elections 2015) gathered from the most popular bloggers are "Twitter " as Datasets. The stages of this study are described in Figure 1.

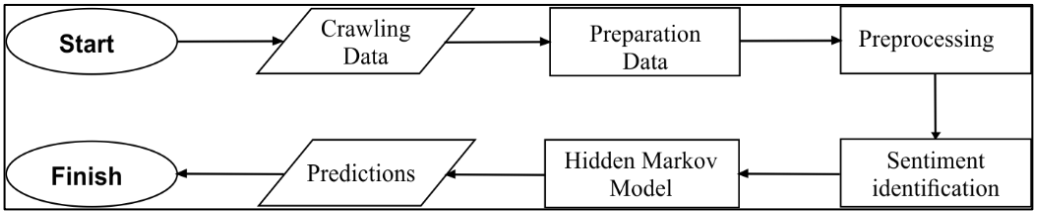

Figure 1. Research algorithm

Based on Figure 1, the first step is crawling data. In this step, data is gathered from the most famous micro blogger site i.e. "Twitter ". The number of tweets is too big so it is impossible to select a manual tweet therefore Python is used "Twitterscraper" as the interface to extract tweets directly from Twitter. Extracted data is captured to a text file in XLSX format (Excel) because it is a human readable format as well as the machine also easy to reduce it. The second is data preparation, data preparation serves to manipulate the data so it looks neat and any variables are needed to be analyzed. In this case the data obtained has 21 variables, but the variable that we analysis just a text variable and a new variable that is a candidate variable.

A text variable is a comment to a candidate while the candidate variable is the name of the candidate that gets the comment. The next is preprocessing data, this is the important steps for data mining processes. The data used in the mining process is not always in the ideal conditions for processing. Sometimes in this data there are a variety of issues that can interfere with the results of the mining process itself such as those with missing values, redundant data, outliers, or data formats incompatible with the system. Therefore, to address this issue required stage preprocessing. Preprocessing is one of the steps of eliminating problems that can interfere with results rather than processing data. In terms of document classification using the type of text data, there are several types of processes that generally include folding case, filtering (removing punctuation), stop word, stemming. The preprocessing stage is as follows:

1. Stop word: a stage for removing unnecessary words such as "yang", di", "ke" and so on. This is done to improve the effectiveness of the system so that the data to be processed is considered important text only. The Stop word used in Python is Sastrawi.
2. Cleaning: a process to clear the document of unnecessary words to reduce the noise in the analysis process.
3. Stemming: a method for mapping the token to its basic form (Rizqon DKK,2017). This is done to change the word that is to be said to be a basic word such as "melepas" to "lepas", "berjumpa" to "jumpa", and so on.

The fourth step is sentiment identification. Identifying the tweet expressions are important thing to do a sentiment analysis with the Python programming language to distinguish the extracted tweets in categories such as positive, negative & neutral, because the extensive library of each word is extracted compared to the popular positive words and negative words. After the classification stage, the tweets score is defined to rely on the complete tweets specified as positive, negative or neutral. There are a number of methods to calculate the sentence sentiment value but here we use one of the popular methods.

$$Sentiment = \frac{P - N}{P + N + O}$$

where, P is positive word, N is Negative Word and O is total words. In this case, we can specify the sentiment. If the value of sentiment is more than 0 then it can be deduced from the positive sentiment. If the value of sentiment is less than 0 it can be deduced from the negative sentiment. If the value of sentiment equals 0, it can be deduced from the Neutral sentiment. The next step is model formation using HMM then the last is prediction the result analysis.

Hidden Markov Model (HMM)

According to [10] a Markov Chain is useful when we need to calculate the probability for an observable sequence of events. However, in HMM, the events we observe are hidden (we don't observe them directly). HMM is formed from several variables i.e. S is the number of states in a Markov model, A is the probability of a state transition, B is the probability of emissions in a state, and π is the initials probability of the state on a Markov model. HMM can be defined as follows:

$$\lambda = (A, B, \pi) \quad (1)$$

A is a probability of transition from state i to State j:

$$A = [P_{ij}],\ P_{ij} = (q_n = s_j \mid q_{n-1} = s_i) \quad (2)$$

B is a probability of emission or likelihood observation which is the probability of O_n:

$$B = [b(k)],\ b_i(k) = P\,(v_n = v_k \mid q_n = s_i) \quad (3)$$

π is an initial probability:

$$\pi = [\pi_i],\ \pi_i = (q_i = s_i) \quad (4)$$

Viterbi Algorithm

According to [11] the Viterbi algorithm aims to find the optimal estimate for the hidden state sequence within HMM, conditional on a series of system measurements. At each stage, the Viterbi algorithm finds the optimal value for the state in the order, and continues the analysis to the next stage in the inductive way. To find the optimal order in the hidden state $Q = (q_1, q_2, \dots\ q_n)$ in the realization of HMM, conditional on the measurement sequence system $O = (o_1, o_2, \dots o_n)$, the following variables are defined:

$$v(j) = \max_{Q} p\,(Q = j \mid \lambda) \tag{5}$$

where $v(j)$ is the optimal value for HMM at the time n, considering the first state of S_i as the condition. The value of $_n(j)$ is calculated as follows:

$$v(j) = \max_{i=1} v_{n-1}(i) P_{ij} b_j(o_n) \tag{6}$$

Then to obtain the best value of P calculated using following formula

$$P * = \max_{i=1} v(i) \tag{7}$$

The three factors are multiplied in equation (6) to extend the previous path by calculating the Viterbi probability at time n are follows:

1. $v_{n-1}(i)$ is the probability of the previous Viterbi path from the previous time step.
2. P_{ij} is the transition probability from the previous state to the current state q_j.
3. $b_j(o_n)$ the observation state against the observation symbol given the current state j

Result and Discussion

This research method uses the Indonesian tweet data taken from 8 May 2015 until 20 September 2015. The amount of data collected is as much 1562 data. After the data is collected, prepared and then preprocessed, then sentiment identification is carried out. In the HMM process, the weight of each word is needed with the model created. The steps in HMM are as follows:

1. Initial state serves to analyze the initial probability. In this study the results of the initial state π are as follows

$$\pi = \begin{bmatrix} \frac{total\ positif}{total\ sentimen} \\ \frac{total\ negatif}{total\ sentimen} \\ \frac{total\ netral}{total\ sentimen} \end{bmatrix} = \begin{bmatrix} \frac{472}{1552} \\ \frac{531}{1552} \\ \frac{549}{1552} \end{bmatrix} = \begin{bmatrix} 0.3042 \\ 0.3421 \\ 0.3537 \end{bmatrix}$$

2. The transition probability in HMM functions to analyze the movement of a probability from one state to another. The transition probability in this study is

$$A = \begin{bmatrix} 0.6059 & 0.2055 & 0.1886 \\ 0.1623 & 0.6830 & 0.1547 \\ 0.1843 & 0.1277 & 0.6880 \end{bmatrix}$$

3. The emission probability in the HMM serves to analyze an observation in the state. emission probability in this study is as follows

$$B = \begin{bmatrix} 0.73728814 & 0.26271186 \\ 0.71186441 & 0.28813559 \\ 0.74087591 & 0.25912409 \end{bmatrix}$$

The calculation above illustrates the transition probability, emission and initial state, a chart showing all the states that occur is shown in the Figure 2.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

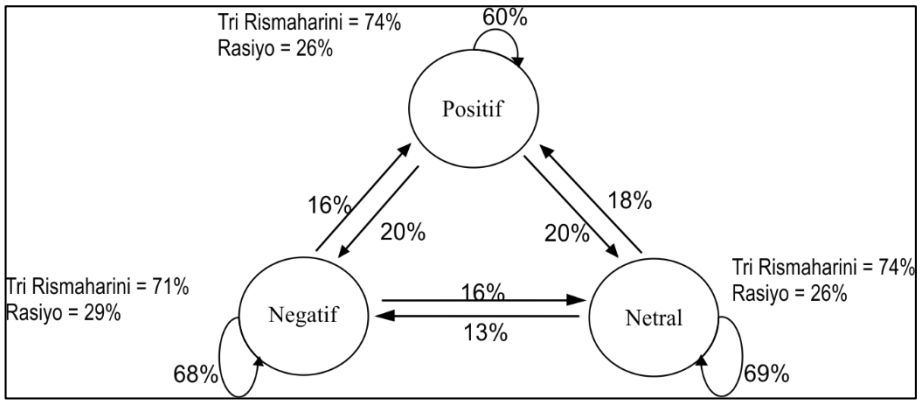

Figure 2. Hidden Markov Model based on transition probability, emission and initial state

Figure 2 can be concluded that with the data obtained via twitter, the candidate's observation of positive sentiment has a 74% percentage for the Tri Rismaharini candidate and 26% for the Rasiyo candidate, the candidate's observation of negative sentiment has a 71% percentage for the Sri Rismaharini candidate and 26% for the Rasiyo candidate, and Candidate observations on neutral sentiment have a percentage of 74% for candidate Tri Rismaharini and 26% for candidate Rasiyo. So, most of the candidates who have the highest percentage of each sentiment are the Tri Rismaharini candidates. However, using the chart above and Markov's assumptions, researchers can easily predict whether the next tweet will be positive, negative or neutral. By using the Viterbi algorithm, researchers can provide the best route for the state in the order of the first day is Tri Rismaharini and the second day is Rasiyo. Probabilities for the observation of Tri Rismaharini and Rasiyo on Hidden State are as follows:

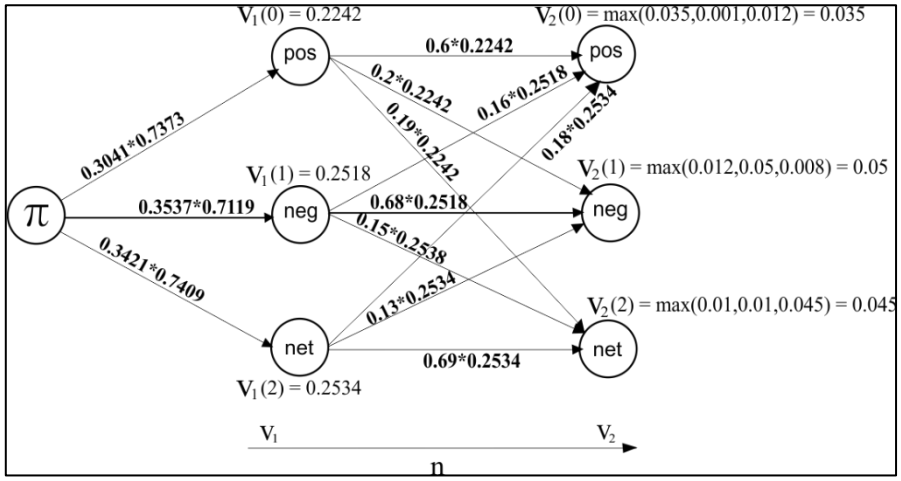

Figure 3. Probabilities for the observation using Viterbi Algorithm

From Figure 3, It can be concluded that the best route is observation on Tri Rismaharini with the prediction of neutral sentiment. While observing Rasiyo with negative sentiment predictions.

Conclusions

The use of HMM to predict produces initial state, transition probability, and emission probability so that it considers hidden states that can affect accurate forecasting. The proposed model is more accurate for predicting sentiment on candidates from the 2015 Pilkada. It also helps future Pilkada to know the candidates based on sentiment so that they can improve each candidate's performance to get good sentiment or they can manage wide publicity to know every sentiment on the candidates. In this case the algorithm of Viterbi has predicted the best route with the candidate Tri Rismaharini gained a prediction of neutral sentiments, whereas Rasiyo candidates gained sentiment negative predictions as well.

Acknowledgements

The authors would like to financial fund support from Ahmad Dahlan University Project fund.

References

[1] F.Ari, Konsep Data Mining.Universitas Jendral Soedirman. Purwokerto. Indonesia (2011).
[2] P. M. Eko, Teori Dasar Hidden Markov Model. Bandung: Institut Teknologi Bandung (2010)
[3] E. Asgarian, M. Kahani, and S, Sharifi, The Impact of Sentiment Features on the Sentiment Polarity Classification in Persian Reviews, Cogn Comput 10 (2018) 117–135.
[4] L. Biu, Sentiment Analysis and Opinion Mining, Department of Computer Science, Chicago (2012)
[5] F. Xing and Z. Justin, Sentiment Analysis using Product Review Data, North California and State University. USA (2015).
[6] S. Gupta and S. K. Gupta, Natural language processing in mining unstructured data from software repositories: a review, Sådhanå, 44(244) (2019) 1-17.
[7] S. Rizqon, Perancangan Sistem Analisis Sentimen Masyarakat Pada Sosial Media dan Portal Berita, Yogyakarta: STMIK AMIKOM Yogyakarta. (2017)
[8] R. Chowdhury, P. Ganapathi, V. Pradhan, J. J. Tithi, and Y. Xiao, An Efficient Cache-oblivious Parallel Viterbi Algorithm, European Conference on Parallel Processing, (2016) 574-587
[9] P. Li and Q.H. Zhang. IF Estimation of Overlapped Multicomponent Signals Based on Viterbi Algorithm, Circuits, Systems, and Signal Processing, 39 (2020) 3105–3124.
[10] J. Daniel, and H. M James, Speech and Language Processing, California: University (2019)
[11] M. Orchard, C.M. Poblete, J.I. Huircan, P. Galeas, and H. Rozas, Harvest Stage Recognition and Potential Fruit Damage Indicator for Berries Based on Hidden Markov Models and the Viterbi Algorithm, Sensors, 19 (20) (2019) 1-16.