hasil cek_Hidden Markov Model for Sentiment Analysis using Algorithm Viterbi

by Sugiyarto Sugiyarto

Submission date: 05-Nov-2020 11:44AM (UTC+0700)
Submission ID: 1436634831
File name: Markov_Model_for_Sentiment_Analysis_using_Algorithm_Viterbi.pdf (278.09K)
Word count: 2343
Character count: 12200

Hidden Markov Model for Sentiment Analysis using Algorithm Viterbi

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

[1,2,3]Department of Mathematic, Faculty of Science and Technology, Ahmad Dahlan University, Indonesia
[4]Department of Elementary School Education, Faculty of Teacher Training and Education, Ahmad Dahlan University, Indonesia

[*] haswat21@gmail.com, sugiyarto@math.uad.ac.id

Abstract: Data mining is an activity to extract or mine knowledge from large amounts of data, this information that will later be used for development. Text becomes very important in some applications, such as the processing and the conclusion of a person's review, and the book, then analysis of opinion in politics, as well as classifying the object. These types of online data have some flaws that could potentially inhibit the sentiment analysis process. The first drawback is that people can freely post their own content, the quality of their opinions cannot be guaranteed. For example, sharing opinion related to topics, online spammers posting spam on forums. Some spam does not mean at all, while others have irrelevant opinions also known as false opinions. The second drawback is that the basic truth of such online data is not always available. Basic truth is more like a tag of a particular opinion, indicating whether the opinion is positive, negative, or neutral. The main objective of this study is to increase the forecasting strength of candidates in Surabaya considering the hidden problems that affect to make the predictions go wrong. This work is done with the help of statistics model namely Hidden Markov Model (HMM) applied to the dataset extracted sentiment at the 2015 elections in Surabaya from the popular site micro blogging "Twitter ".

Keywords: Hidden Markov Model, Stochastic, Sentiment Analysis

Introduction

Data mining is an activity to extract or mine knowledge from large amounts of data, this information that will later be used for development. 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 [4].

Hidden Markov Model (HMM) is a statistical model where a system is in it by modeling as Markov processes on 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[8,14].

Sentiment analysis plays an important role to classify data into positive, negative, and neutral 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 [7]. 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. Text becomes very important in some applications, such as the processing and the conclusion of a person's review, and the book, then analysis of opinion in politics, as well as classifying the object. These types of online data have some flaws that could potentially inhibit the sentiment analysis process. The first drawback is that people can freely post their own content, the quality of their opinions cannot be guaranteed. For example, sharing opinion related to topics, online spammers posting spam on forums. Some spam does not mean at all, while others have irrelevant opinions also known as false opinions. The second drawback is that the basic truth of such online data is not always available. The basic truth is more like a tag of a particular opinion, indicating whether the opinion is positive, negative, or neutral [5].

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 [11]. The source of the data is very large, one example is Twitter that contains about the elections 2020 that will be implemented simultaneously in various regions. But the data is very much and large, humans are not able to group one. Designing a system that is able to analyse the sentiment of society on social media as one solution to solve the problem of election 2015. So sentiment analysis is needed to analyze large, unstructured data such as through Twitter about the 2015 elections

Materials and Methods

Researchers use 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.

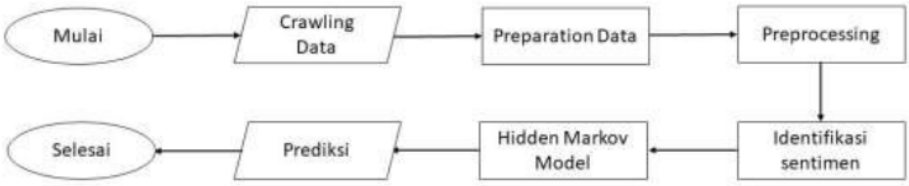

Figure 1 Research algorithm

Crawling data

In this step, data is gathered from the most famous microblogger site i.e. "Twitter ". The number of tweets is too big so it is impossible to select a manual tweets, therefore Python is used with "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.

Preparaton data

Preparation data 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 Cuman 1 is a text variable and plus 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.

Preprocessing

Preprocessing is one of 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), stopword, stemming. The preprocessing stage is as follows:

1. Stopword
   Stopwords are 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 Stopword used in Python is Sastrawi.
2. Cleaning
   Cleaning is a process to clear the document of unnecessary words to reduce the noise in the analysis process.
3. Stemming
   Stemming is 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.

Identifikastion Sentiment

Identifying the expression of tweets we need 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 = Positive Word
N = Negative Word
O = 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 sentiment positively. If the value of sentiment is less than 0 it can be deduced from the sentiment Negtaif. If the value of sentiment equals 0, it can be deduced from the sentiment Neutral.

Hidden Markov Model

According to Daniel Jurafsky & James H. Martin (2019). 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 state 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 probability of initials 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:

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

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

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

π is an early probability:

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

1. Viterbi Algorithm

   According to the Marcos Orchard at.al (2019), 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:

   $$v_n(j) = \max_Q p(Q = j | \lambda) \quad (5)$$

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

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

   To obtain the best value of the Dihitiung as follows:

   $$P* = \max_{i=1} v_N(i) \quad (7)$$

Result and Discussion

- The Model proposed above, the following probability calculated for the matrix of probability transitions from HMM are:
  $P(q_n = Positif | q_{n-1} = Positif) = 0.6$
  $P(q_n = Positif | q_{n-1} = Negatif) = 0.2$
  $P(q_n = Positif | q_{n-1} = Netral) = 0.2$
  $P(q_n = Negatif | q_{n-1} = Positif) = 0.16$
  $P(q_n = Negatif | q_{n-1} = Negatif) = 0.68$
  $P(q_n = Negatif | q_{n-1} = Netral) = 0.16$
  $P(q_n = Netral | q_{n-1} = Positif) = 0.18$
  $P(q_n = Netral | q_{n-1} = Negatif) = 0.13$
  $P(q_n = Netral | q_{n-1} = Netral) = 0.69$
- While the probability of emissions from HMM are:
  $P(v_n = Tri\ Rismaharini | q_n = Positif) = 0.74$
  $P(v_n = Tri\ Rismaharini | q_n = Negatif) = 0.71$
  $P(v_n = Tri\ Rismaharini | q_n = Netral) = 0.74$
  $P(v_n = Rasiyo | q_n = Positif) = 0.26$
  $P(v_n = Rasiyo | q_n = Negatif) = 0.29$
  $P(v_n = Rasiyo | q_n = Netral) = 0.26$

The following table illustrates the probability of transitions and emissions, the circle in the chart signifies all possible state.

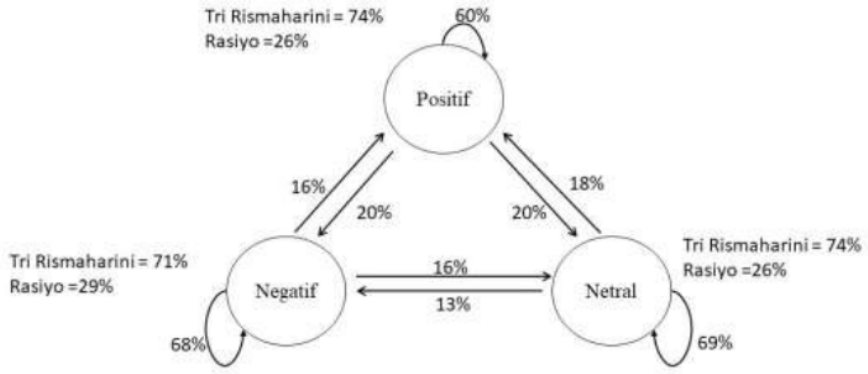

Figure 2 Hidden Markov Model

However, using the above chart and Markov's assumption, researchers can easily predict whether subsequent tweets will be positive, negative or neutral. By using the algorithm of viterbi, researchers can provide the best route for the state in the order of Tri Rismaharini and Rasiyo. Here are the probability calculations on the viterbi algorithm:

1. Probabilities for the observation of Tri Rismaharini and Rasiyo on Hidden State:

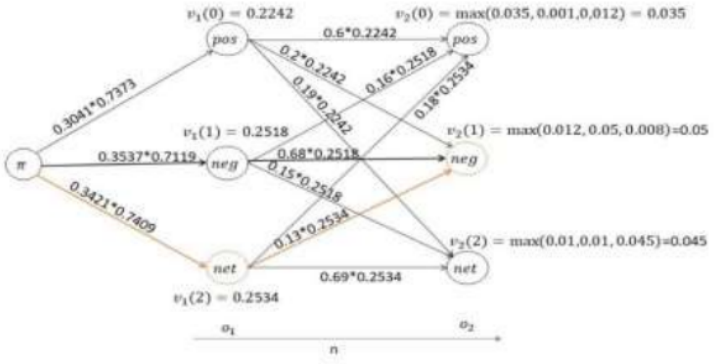

Figure 3 Viterbi Algorithm

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

With 1562 data analyzed as a dataset. Hidden Markov Model to consider the hidden state that can affect accurate forecasting. The proposed Model is more accurate to predict candidate features. It also helps politics to introduce candidates based on reviews so that they can increase candidate performance or they can manage broad publicity to promote candidates. In 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] Aggarwal, C.,Charu.Data Mining:The Textbook.new York:Springer Cham Heidelberg.2015
[2] Aporvm, et al. Sentiment Analysis of Twitter Data.Columbia University. New York.2015
[3] Feldman, R.,Sanger, J. The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data. New York : Cambridge University Press.2007
[4] Fadli.Ari.. Konsep Data Mining.Universitas Jendral Soedirman. Purwokerto. Indonesia.2011
[5] Fang,Xing.,Zhan,Justin.Sentiment Analysis using Product Review Data.North California A and T State University.USA.2015
[6] Jurafsky,Daniel.,H.Martin,James. Speech and Language Processing. University. California. 2019
[7] Liu,Biu. Sentiment Analysis and Opinion Mining. Department of Computer Science.Chicago.2012
[8] Prasetyo.Muhammad, Eko. Teori Dasar Hidden Markov Model. Institut Teknologi Bandung. Bandung. 2010
[9] Ross,Sheldon M. Introduction Probability Models.University of Southern California.Los Angeles,California. 2010
[10] Saputra,Nurirwan.,Teguh,Bharata A., Erna P, Adhistya. Analisis Sentimen Data Presiden Jokowi dengan Preprocessing Normalisasi dan Stemming Menggunakan Metode Naive Bayes dan SVM. Universitas PGRI Yogyakarta.Yogyakarta.2015
[11] Sadida,Rizqon., dkk. Perancangan Sistem Analisis Sentimen Masyarakat Pada Sosial Media dan Portal Berita. STMIK AMIKOM Yogyakarta. Yogyakarta. 2017
[12] Santoso,Budi, Suharyanto, Djoko Legono. Penerapan Optimasi Parameter pada Metode Exponential Smoothing untuk Perkiraan Debit.Universitas Gunadarma, Univeritas Diponegoro, Universitas Gadjah Mada. Indonesia. 2009

[13] Orchard,Markos.,dkk.(2019). Harvest Stage Recognition and Potential Fruit Damage Indicator for Berries Based on Hidden Markov Models and the Viterbi Algorithm.
[14] Universitas perbatasan Chili.Chili.Walpole,Ronald E.(1997).Pengantar Statistika edisi ke-4.Jakarta: PT. Gramedia Pustaka Utama
[15] Abdul Raffey,Mohd, et al. Forecasting Product Sale from Twitter using Hidden Markov Model.Aurangabad, india. 2018
[16] Rima A.,Aditya. Penggunaan Web crawler untuk menghimpun tweets dengan metode pre-processing Text Mining.Universitas Telkom, Indonesia. 2015

hasil cek_Hidden Markov Model for Sentiment Analysis using Algorithm Viterbi

ORIGINALITY REPORT

13%	11%	11%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journalofbigdata.springeropen.com Internet Source	8%
2	www.mdpi.com Internet Source	1%
3	mafiadoc.com Internet Source	1%
4	Marcos Orchard, Carlos Muñoz-Poblete, Juan Ignacio Huircan, Patricio Galeas, Heraldo Rozas. "Harvest Stage Recognition and Potential Fruit Damage Indicator for Berries Based on Hidden Markov Models and the Viterbi Algorithm", Sensors, 2019 Publication	1%
5	Evaristus Didik Madyatmadja, Debri Pristinella, Martinus Damitutsa Kurnia Dewa, Hendro Nindito, Cristofer Wijaya. "Data Mining Techniques of Complaint Reports for E-government: A Systematic Literature Review", 2020 International Conference on Information	<1%

Management and Technology (ICIMTech), 2020
Publication

6	www.journal.uad.ac.id Internet Source	<1%
7	Winda Widya Ariestya, Ida Astuti, I Made Wiryana. "Preprocessing For Crawler Of Short Message Social Media", 2018 Third International Conference on Informatics and Computing (ICIC), 2018 Publication	<1%
8	journals.sagepub.com Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		