JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 5, No. 3. June 2023 DOI: https://doi.org/10.34288/jri.v5i3.227

Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Latent Dirichlet Allocation for Uncovering Fraud Cases on Twitter

Sallu Muharomah-[1], Chanifah Indah Ratnasari-[2*)]

[1,2] Informatics
Universitas Islam Indonesia
Yogyakarta, Indonesia
sallu.muharomah@students.uii.ac.id[1], chanifah.indah@uii.ac.id[2]

(*) Corresponding Author

Abstract

Fraud is a phenomenon that continues to exist in society with a modus operandi that continues to evolve with the times. The mode of operation of fraud is continually evolving with technological advancements, globalization, and consumer behavior shifts. In today's digital age, social media is important in spreading information regarding fraud. Twitter is a social media platform that is widely used. Twitter provides easy and fast access to relevant information. As a result, to raise fraud awareness, it is critical to study the mode of operation of fraud spread on social media, particularly on Twitter. The Latent Dirichlet Allocation (LDA) approach is used in this work to classify and identify fraud issues often addressed by Indonesian Twitter users. By applying LDA modeling, this study aims to understand more comprehensively the fraudulent topics that often appear on Twitter. The research found that seven fraud topics are most commonly discussed by Twitter users in Indonesia, with the highest cohesion value of 0.491899.

Keywords: Fraud; Topic Modeling; Twitter; LDA

Abstrak

Penipuan merupakan fenomena yang terus eksis dalam masyarakat dengan modus operandi yang terus berkembang seiring perkembangan zaman. Modus operandi penipuan senantiasa berubah dan berkembang seiring dengan kemajuan teknologi, globalisasi, dan pergeseran perilaku konsumen. Dalam era digital saat ini, media sosial memiliki peran signifikan dalam penyebaran informasi mengenai penipuan. Twitter, sebagai salah satu platform media sosial yang banyak digunakan. Twitter memberikan akses mudah dan cepat terhadap informasi yang relevan. Oleh karena itu, penelitian tentang modus operandi penipuan yang tersebar di media sosial, terutama di Twitter, penting untuk dilakukan. Penelitian ini melakukan pemodelan topik dengan menggunakan metode Latent Dirichlet Allocation (LDA) untuk mengklasifikasikan dan mengidentifikasi topik-topik penipuan yang sering dibahas oleh pengguna Twitter di Indonesia. Dengan menerapkan pemodelan LDA, penelitian ini bertujuan untuk memahami lebih komprehensif mengenai topik-topik penipuan yang sering muncul di Twitter. Berdasarkan penelitian yang dilakukan, ditemukan bahwa terdapat 7 topik penipuan yang paling umum dibahas oleh pengguna Twitter di Indonesia dengan nilai kohesi tertinggi sebesar 0.491899.

Kata kunci: Penipuan; Pemodelan Topik; Twitter; LDA

INTRODUCTION

The practice of fraud is a phenomenon that always exists in society, with continuous changes in the modus operandi adapted to the times (Kurnia et al., 2022). Fraud itself is an act that involves falsifying, manipulating, or deceiving information by entering into illegal benefits from other people (Yuwita et al., 2022). The modus operandi of fraud is constantly changing and developing, along with technological advances, globalization, and shifts in consumer behavior (Wahyudi et al., 2022). Fraudsters use clever methods and disguise their identities, making it difficult for potential victims to identify and deal with them (Puspitasari, 2018). Therefore, it is imperative to know about the modus operandi and indications of fraud so that individuals can anticipate and avoid being trapped in harmful fraudulent practices.

In the current digital era, the role of social media is increasingly significant in disseminating information about fraud (Nur, 2021). Social media, such as Twitter, has become a platform that is widely used by the public to obtain the latest information (Saura et al., 2019), including information regarding the modus operandi of fraud (Trinugraheni, 2022), (Khatulistiwa, 2021), and (Shalihah, 2021). Twitter allows individuals to share experiences and stories and warn others about potential fraud that is currently circulating (Hafis, 2020). Thus, Twitter can provide easier and faster access to information about fraud so that individuals can be more vigilant and take appropriate preventive measures.

Therefore, in-depth research on the modus operandi of fraud spread on social media is essential. Collective efforts in educating the public, raising awareness of fraud tactics, and promoting digital literacy can lower the success rate of fraud and shield people from the financial and emotional harm that fraudulent practices cause. Therefore, a more structured approach is needed. A topic modeling approach could be a viable answer. This initial research will analyze fraud cases that occurred on Twitter in 2022.

This research aims to classify and identify the most common fraud issues Indonesian Twitter users discuss. This study employs Latent Dirichlet Allocation (LDA) modeling to gain a better understanding of the fraud-related topics that Indonesian Twitter users frequently discuss. A thorough understanding of these areas can help anticipate and avert more sophisticated, potentially costly fraud instances in the future.

RESEARCH METHODS

In this study, seven stages have been carried out, as shown in Figure 1.

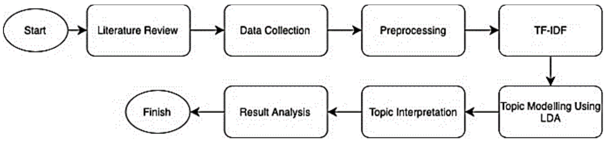

Figure 1. Research Stage

The first stage consists of a literature review, followed by data collection, preprocessing, TF-IDF weighting, the topic modeling process using the LDA method, topic interpretation, and finally, the outcome analysis.

Literature Review

Before picking the approach, this study conducted a literature review. A method is required to help steer the process of implementing topic modeling.

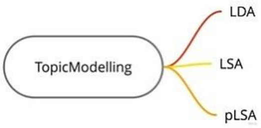

Figure 2. Kinds Of Topic Modeling Methods

Methods can be used to increase the quality of topic modeling results for better analysis or decision-making. Figure 2 depicts three types of topic modeling often utilized (Mandal, 2020). In this part, three past research topics' models are used to compare each system development method.

First, research by Nurlayli and Nasichuddin (2019) entitled "Topic Modeling Penelitian Dosen JPTEI UNY pada Google Scholar Menggunakan Latent Dirichlet Allocation." This paper covers modeling research themes based on Google Scholar publication data from various JPTEI UNY lecturers to determine research trends undertaken by lecturers in a department or study program. The LDA approach is utilized to aid in discovering research topics that JPTEI UNY lecturers frequently address.

Second, the study "Twitter and market efficiency in the energy market: Evidence using LDA cluster topic extraction" by Polyzos and Wang (2022). This study used tweet samples to evaluate and measure market energy efficiency. Using the Latent Dirichlet Allocation technique to collect the topics that arise on days when the market index rises and falls, it was discovered that the topics of tweets on days of ups and downs differ.

Third, research by Gupta and Patel (2021) entitled "Method of Text Summarization Using LSA And Sentence-Based Topic Modeling with Bert." This research executes essential NLP tasks, such as summarizing documents and modeling topics. This study aims to provide a short and fluid summary of long text material while keeping relevant information in the document. It employs LSA as a way of extracting relevant topics from text documents.

Furthermore, research has yet to use the PLSA approach for topic modeling. As a result, the Latent Dirichlet Allocation (LDA) method will be used in this study. LDA is a generative probabilistic model that processes discrete data like text (Kannitha et al., 2022). This model represents documents as a random mix of latent (invisible) topics (Bhat et al., 2020). LDA is a three-level hierarchical Bayesian model where each topic is modeled as an infinite mix through a set of underlying topic probabilities, and each collection item is described as a finite mix of a set of topic sets. (Suhartono, 2018).

Preprocessing

Data gathered from the Twitter scraping process is preprocessed to ensure the dataset's integrity, quality, and usability. These preprocessing techniques handle various issues,

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v5i3.227
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

including data loss, excessive data, and inconsistencies. Figure 3 depicts the steps.

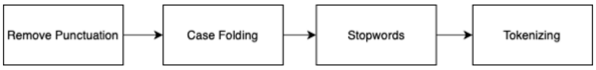

Figure 3. Preprocessing steps

Preprocessing in this study consists of multiple phases, including case folding, punctuation removal, stopwords, and tokenizing (Kaila, 2020). Case folding is the first stage in the preprocessing process. This stage will alter the uppercase to the lowercase character of the letters in the data. The goal of this stage is to avoid reading the same word twice but being misunderstood due to differences in upper and lower case (Kaila, 2020).

After case folding, the next step is to remove punctuation. This stage eliminates unneeded characters like numerals, punctuation marks, links or tags, and special characters (%, $, &, etc.) (Kaila, 2020). The next stage is stopwords. A stoplist, including irrelevant or insignificant terms, will be prepared at this step. This stage filters unnecessary terms, with the remaining words regarded as important or keywords (Kaila, 2020). The last stage of preprocessing is tokenizing. Tokenizing is a step that separates words in data sentences to make it easier to calculate words to transform the data in the following step of the study (Chilmi, 2021).

TF-IDF

A numerical statistic called TF-IDF (Term Frequency-Inverse Document Frequency) is used to assess the significance of a term in a document or group of documents. It combines two components, term frequency (TF) and inverse document frequency (IDF) (Sasmita & Falani, 2018).

Term Frequency (TF) measures how frequently a term appears in a document. It is determined by dividing a document's total number of terms by the frequency with which each time appears. The theory underlying TF is that a period may be more significant in describing the content of a document if it occurs more frequently in that document (Curiskis et al., 2020).

A term's rarity or uniqueness throughout the document collection is gauged via the Inverse Document Frequency (IDF) method. The ratio between the total number of documents in the collection and the number of documents containing the phrase is calculated as its logarithm. The IDF component helps down-weight phrases frequently used in papers because they are less helpful in identifying one document from another (Silveira et al., 2021).

The TF-IDF score for a term in a document is obtained by multiplying its TF by its IDF. The resulting score reflects the relative importance of the terms within the document and across the collection. Higher TF-IDF scores indicate that a term is more significant to a particular document or set of documents (Zhou et al., 2020).

Latent Dirichlet Allocation (LDA)

After preprocessing, the LDA algorithm is used to model the topics in the document collection. LDA, or Latent Dirichlet Allocation, is used in text analysis to identify and classify hidden topics in a collection of documents (Fahlevvi & SN, 2022). Each document is considered a combination of several hidden topics in this process. LDA operates under the assumption that every word in the document originates from one of the topics listed.

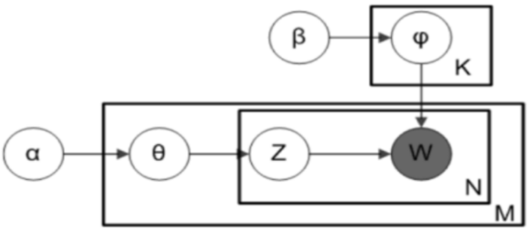

Figure 4. LDA Graphical Model

In Latent Dirichlet Allocation (LDA), a popular generative probabilistic model for topic modeling, a graphical model representation can be depicted using plate notation, as shown in Figure 4. Plate notation is a visual way of representing repeated variables or entities in a probabilistic model. The variables in the plate notation for LDA are as follows:

β: The Dirichlet parameter for the distribution of words on topics. It controls the topic-word distribution. β is not shown explicitly in the plate notation but is used to generate the distribution φ.

φ: The distribution of words to the topics in the corpus. It represents the topic-word distribution. φ is generated from the Dirichlet distribution with parameter β.

K: A collection of topics. K represents the total number of topics in the model. Each topic is associated with a distribution of words.

W: A word. W represents an observed word in the corpus.

N: A group of words. N represents the total number of words in a document. Each word in the document is associated with a topic assignment.

M: A document set. M represents the total number of documents in the corpus. Each document consists of a collection of words.

Z: The index assignment of a topic. Z represents the hidden topic assignment for each word in the document.

θ: Document topic distribution. θ represents the distribution of topics in a document. It generates from the Dirichlet distribution with parameter α.

α: The Dirichlet parameter for the distribution of topics to documents. It controls the document-topic distribution. α is not shown explicitly in the plate notation but is used to generate the distribution θ.

Overall, the plate notation for LDA visually represents how the different variables and their dependencies relate to each other in the generative process of the model. It helps understand the flow of the probabilistic model and the relationships between the variables involved. LDA is formulated in equation (1) (Zvornicanin, 2022).

$$\mathrm{p(w, z, \theta, \varphi|\alpha, \beta)} = p(\varphi|\beta)p(\theta|\alpha)p(z|\theta)p(w|\varphi k) \text{ ……………………..}(1)$$

At this stage, it provides an understanding of the topics generated by the LDA algorithm. After the topic modeling process, the result is a word probability distribution for each topic. On-topic interpretation involves identifying and naming topics based on the most relevant words with a high probability in each topic.

RESULTS AND DISCUSSION

Scraping Twitter data from January 1, 2022, to December 31, 2022, was performed using Twint with the query "penipuan" (fraud), resulting in 11,581 tweets in CSV format. The collected data is then preprocessed to prevent data value loss, extra data, and inconsistencies in the scraped data. Preprocessing is performed in steps, beginning with the removal of punctuation, followed by case folding, the removal of stopwords from the link https://bit.ly/stopwordss, and finally, the dividing of the text into tokens or smaller pieces such as words, phrases, or sentences. Table 1 shows the sample data that has been processed, and the data will be handled in the next stage.

Topic Interpretation

Table 1. Tweet Data After Preprocessing

Tweet	Tweet After Preprocessing
Kembalikan duit kami !!, Konser berkedok penipuan @msklg_ - Tandatangani Petisi! https://t.co/pUNi452ete via @ChangeOrg_ID	['kembalikan', 'duit', 'konser', 'tandatangani', 'petisi', 'via']
Kronologi? Kaya gini sih, hati hati buat temen temen yang lagi nyari kerja, jangan terlena sama platform platform lowongan pekerjaan yang ga bisa dipertanggung jawabkan gini. Ga semua platform, tapi banyak penipuan kek gini. https://t.co/C6WLaoOueJ	['kronologi', 'kaya', 'temen', 'temen', 'nyari', 'kerja', 'terlena', 'platform', 'platform', 'lowongan', 'pekerjan', 'dipertangung', 'jawabkan', 'platform', 'kek']

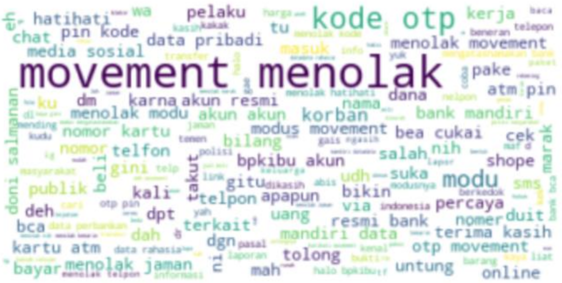

Figure 5. Word Cloud Tweet's Data "Penipuan"

Figure 5 shows the word cloud of "penipuan" tweet data. The word cloud has a function to display words that appear frequently in the dataset. If the writing on the word is getting bigger, the word appears the most, and vice versa. Based on Figure 4, The words in large size are shown in words "movement," "menolak," "kode," and "OTP."

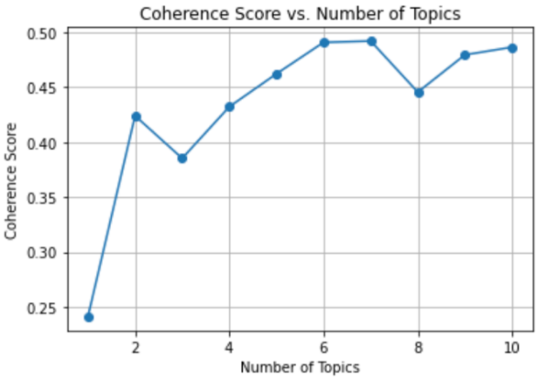

Figure 6. Value-Coherent Graph

The weighting uses id2word.doc2bow to convert text into a TF (Term Frequency) representation. The input data will be broken down into words (tokenization), and then the frequency of occurrence of each word in the document will be calculated. In Figure 6, the highest score is shown on the seventh topic. Table 2 shows the results of word weighting.

Table 2. Topic Coherent Value

Num Topics	Coherence Value
1	0.241203
2	0.42388
3	0.385201
4	0.432102
5	0.462078
6	0.490681
7	**0.491899**
8	0.445436
9	0.479342
10	0.486293

In addition to displaying the visuals of the graphs in Figure 6, Table 2 shows the topic of coherence values. The table proves that the seventh topic has the highest coherence value among other topics, so it is highly relevant in discussing information about fraud. Then it was analyzed using LDA by looking at the visualization of the pyLDAvis tools and the relationship between the words of a topic. The LDA results prove that the seventh topic has 30 terms, which is the most, as shown in Figure 7.

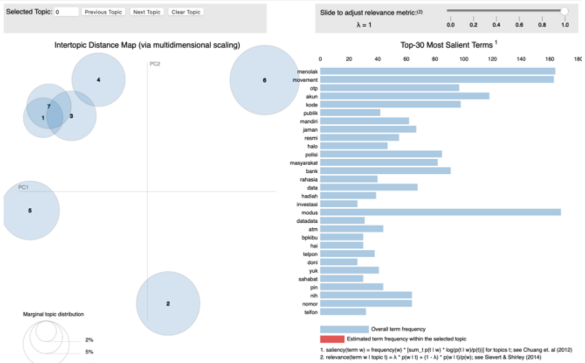
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
6
7
1
3
5
2
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms 1
0
20
40
60
80
100
120
140
160
180
menolak
movement
otp
akun
kode
publik
mandiri
jaman
resmi
halo
polisi
masyarakat
bank
rahasia
data
hadiah
investasi
modus
datadata
atm
bpkibu
hai
telpon
doni
yuk
sahabat
pin
nih
nomor
telfon
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 7. LDA Visualization of “penipuan” Tweets

The location of the bubble coordinates is based on the calculated weight values. If there are bubbles that are scattered and do not intersect, then there is no connection between the topics, and vice versa. Furthermore, it can be seen in Figure 6 that the bubbles collide, so it can be concluded that there are overlapping topics, namely:

- Topic 3 intersects with topics 7, 1, and 4.
- Topic 1 overlaps with Topics 3 and 7.
- Topic 7 intersects with Topics 1 and 3.

Table 3. Topic mapping "penipuan" on Twitter from PC

Quadrant 1	Quadrant 2	Quadrant 4	Quadrant 4
Topic 6	Topic 2	Topic 2	Topic 5
		Topic 5	Topic 4 intersects
			Topic 3 intersects
			Topic 7 intersects
			Topic 1 intersects

Figure 7 shows the bubbles spread across the 4 Principal Components (PC), as shown in Table 3. Each quadrant has its topic, whereas quadrant-4 has the same topic, as seen from the overlapping topics.

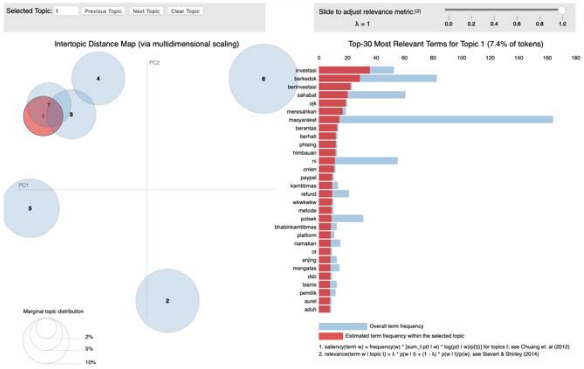
Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
6
7
1
3
5
2
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (7.4% of tokens)
0
20
40
60
80
100
120
140
160
180
investasi
berkedok
berinvestasi
sahabat
ojk
meresahkan
masyarakat
berantas
berhati
phising
himbauan
ni
onlen
paypal
kamtibmas
refund
wkwkwkw
metode
polsek
bhabinkamtibmas
platform
namakan
ot
anjing
mengatas
dsb
bisnis
pemilik
aurel
aduh
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 8. Visualization of Topic-1

Topic-1 is located in quadrant-4. The words that appear are shown in Figure 8. The words that have the highest probability and are representative of Topic-1 are: ‘0.008*"investasi" (investment) + 0.006*"berkedok" (under the guise) + 0.005*"berinvestasi" (invest) + 0.004*"sahabat" (best friend) + 0.004*"ojk" (OJK is an abbreviation of Otoritas Jasa Keuangan: Financial Services Authority) + 0.004*"meresahkan" (unsettling) + 0.003*"masyarakat" (public) + 0.003*"berantas" (eradicate) + 0.003*"berhati" (be careful) + 0.003*"phising”’. And the highest scores were found for the words "investasi" (investment) and "berkedok" (under the guise).

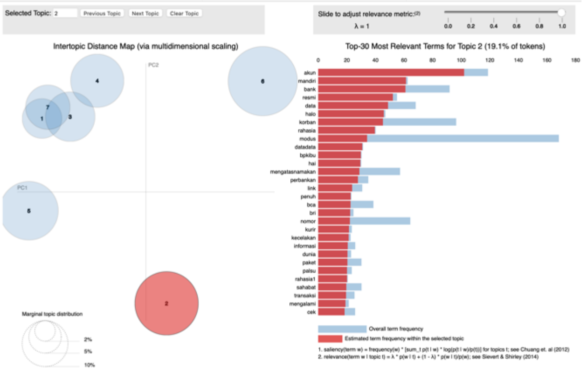

Figure 9. Visualization of Topic-2

Topic-2 is located in quadrant-2 and quadrant-3. The words that appear are shown in Figure 9. The words that have the highest probability and are representative of Topic 2 are: ' 0.018*"akun" (account) + 0.011*"mandiri" (the name of one of the banks in Indonesia) + 0.010*"bank" + 0.009*"resmi" (official) + ''0.008*"data" + 0.008*"halo" (hello) + 0.008*"korban" (victim) + 0.007*"rahasia" (confidential) + ''0.006*"modus" (mode) + 0.005*"datadata"' (data). And the highest scores were found for the words "akun" (account) and "mandiri" (name of bank).

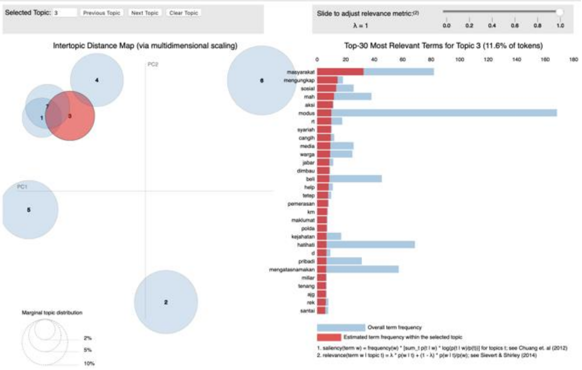

Figure 10. Visualization of Topic-3

Topic-3 is located in quadrant-4. The words that appear are shown in Figure 10, and the words that have the highest probability and are representative of Topic 3 are: '0.009*"masyarakat" (society) + 0.004*"mengungkap" (reveal) + 0.004*"sosial" (social) + 0.003*"mah" (call for mom) + ''0.003*"aksi" (action) + 0.003*"modus" (mode) + 0.003*"rt" (retweet) + 0.003*"syariah" (Sharia) + ''0.003*"cangih" (sophisticated) + 0.003*"media" (media)'. And the highest scores were found for the words "masyarakat," “mengungkap” (reveal), and “sosial" (social).

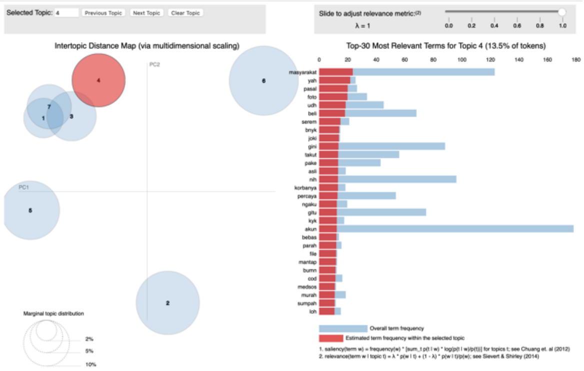

Figure 11. Visualization of Topic-4

Topic-4 is located in quadrant-4. The words that appear are shown in Figure 11. The words that have the highest probability and are representative of Topic-4 are: '0.004*"masyarakat" (society) + 0.004*"yah" (informal language to express disappointment) + 0.003*"pasal" (paragraph) + 0.003*"foto" (photo) + ''0.003*"udh" (done) + 0.003*"beli" (buy) + 0.002*"serem" (scary) + 0.002*"bnyk" (many) + 0.002*"joki" (assistant) + ''0.002*"gini" (thus) '. And the highest scores were found for the words "masyarakat" (society) and “yah" (informal language to express disappointment).

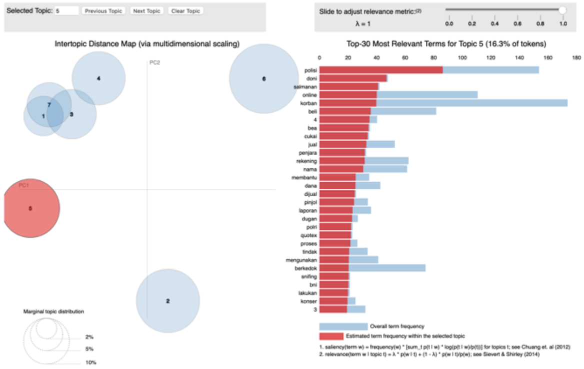

Figure 12. Visualization of Topic-5

Topic-5 is located in quadrant-3 and quadrant-4. The words that appear are shown in Figure 12, and the words that have the highest probability and are representative of Topic-5 are: '0.010*"polisi" (police) + 0.005*"doni" (name of person) + 0.005*"salmanan" (name of person) + 0.004*"online" + ''0.004*"korban" (victim) + 0.004*"beli" (buy) + 0.004*"4" + 0.004*"bea" (tax) + 0.004*"cukai" (tax) + ''0.004*"jual" (sell)'. And the highest scores were found for the words "polisi" (police), “doni” (name of person), and “salmanan" (name of person).

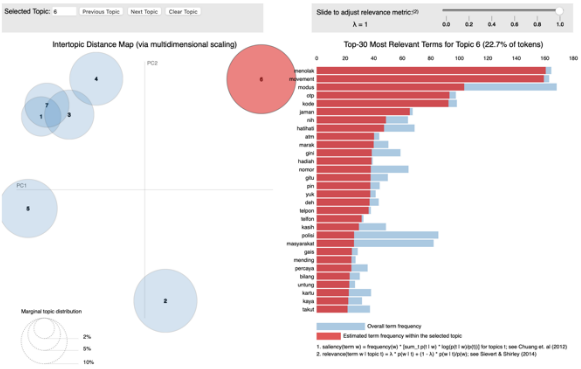
Selected Topic: 6
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 6 (22.7% of tokens)
0
20
40
60
80
100
120
140
160
180
menolak
movement
modus
otp
kode
jaman
nih
hatihati
atm
marak
gini
hadiah
nomor
gitu
pin
yuk
deh
telpon
telfon
kasih
polisi
masyarakat
gais
mending
percaya
bilang
untung
kartu
kaya
takut
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 13. Visualization of Topic-6

Topic-6 is located in quadrant-1. The words that appear are shown in Figure 13. The words that have the highest probability and are representative of Topic-6 are: '0.023*"menolak" (reject) + 0.023*"movement" + 0.015*"modus" (mode) + 0.013*"otp" (on-time password) + ''0.013*"kode" (code) + 0.009*"jaman" (period) + 0.007*"nih" (informal language to express giving) + 0.007*"hatihati" (be careful) + 0.006*"atm" (automated teller machine) ''+ 0.006*"marak" (flare)'. And the highest scores were found for the words "menolak" (reject), “movement”, and “modus" (mode).

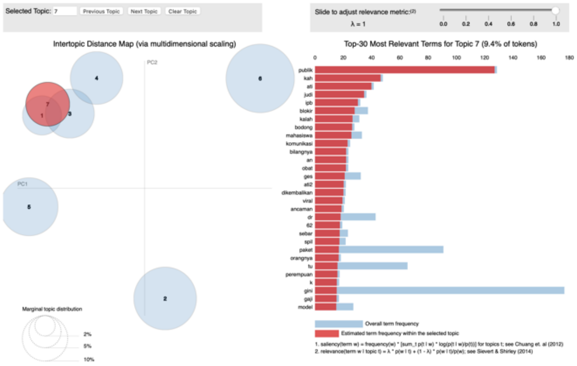
Selected Topic: 7
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 7 (9.4% of tokens)
0
20
40
60
80
100
120
140
160
180
publik
kah
ati
judi
ipb
blokir
kalah
bodong
mahasiswa
komunikasi
bilangnya
an
obat
ges
ati2
dikembalikan
viral
ancaman
dr
62
sebar
spil
paket
orangnya
tu
perempuan
k
gini
gaji
model
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 14. Visualization of Topic-7

Topic-7 is located in quadrant-4. The words that appear are shown in Figure 14. The words that have the highest probability and are representative of Topic-7 are: '0.015*"publik" (public) + 0.005*"kah" (informal language to ask questions) + 0.005*"ati" (be careful) + 0.004*"judi" (gamble) + 0.004*"ipb" (name of institute) + ''0.003*"blokir" (block) + 0.003*"kalah" (lose) + 0.003*"bodong" (bulging) + 0.003*"mahasiswa" (college student) + ''0.003*"komunikasi" (communication)'. And the highest scores were found for the words "publik" (public), “kah” (informal language to ask questions), and “ati" (be careful).

The model for each topic for “penipuan” tweets data is presented in the form of Table 4 as follows:

Table 4. Topic Result Comparison

Topic	Model	Conclusion
6	'0.023*"menolak" (reject) + 0.023*"movement" + 0.015*"modus" (mode) + 0.013*"otp" (on-time password) + ''0.013*"kode" (code) + 0.009*"jaman" (period) + 0.007*"nih" (informal language to express giving) + 0.007*"hatihati" (be careful) + 0.006*"atm" (automated teller machine) ''+ 0.006*"marak" (flare)'	Trend #MovementMenolakPenipuan banyak memberi informasi mengenai masih banyak yang memberikan kode OTP yang mereka miliki kepada penipu (The #MovementMenolakPenipuan trend provides much information about how many people still give their OTP codes to fraudsters)
2	‘ 0.018*"akun" (account) + 0.011*"mandiri" (the name of one of the banks in Indonesia) + 0.010*"bank" + 0.009*"resmi" (official) + ''0.008*"data" + 0.008*"halo" (hello) + 0.008*"korban" (victim) + 0.007*"rahasia" (confidential) + ''0.006*"modus" (mode) + 0.005*"datadata"’ (data)’	Modus mengatasnamakan dari akun Bank Mandiri meminta data-data rahasia nasabah (The mode of acting on behalf of a Bank Mandiri account requests confidential customer data)
5	'0.010*"polisi" (police) + 0.005*"doni" (name of person) + 0.005*"salmanan" (name of person) + 0.004*"online" + ''0.004*"korban" (victim) + 0.004*"beli" (buy) + 0.004*"4" + 0.004*"bea" (tax) + 0.004*"cukai" (tax) + ''0.004*"jual" (sell)'	Polisi mengungkap banyak korban dari doni salmanan (The police uncovered many victims from Doni Salmanan)
4	'0.010*"polisi" (police) + 0.005*"doni" (name of person) + 0.005*"salmanan" (name of person) + 0.004*"online" + ''0.004*"korban" (victim) + 0.004*"beli" (buy) + 0.004*"4" + 0.004*"bea" (tax) + 0.004*"cukai" (tax) + ''0.004*"jual" (sell)'	Penipuan yang dilakukan joki apapun entah itu konser atau joki masuk perguruan tinggi (Fraud any jockey commits, whether it's a concert or a jockey entering college)

Topic	Model	Conclusion
3	'0.009*"masyarakat" (society) + 0.004*"mengungkap" (reveal) + 0.004*"sosial" (social) + 0.003*"mah" (call for mom) + ''0.003*"aksi" (action) + 0.003*"modus" (mode) + 0.003*"rt" (retweet) + 0.003*"syariah" (sharia) + ''0.003*"cangih" (sophisticated) + 0.003*"media" (media)'	Banyak masyarakat yang terlibat penipuan berlabel syariah (Many people are involved in fraud labeled Sharia)
7	'0.015*"publik" (public) + 0.005*"kah" (informal language to ask questions) + 0.005*"ati" (be careful) + 0.004*"judi" (gamble) + 0.004*"ipb" (name of institute) + ''0.003*"blokir" (block) + 0.003*"kalah" (lose) + 0.003*"bodong" (bulging) + 0.003*"mahasiswa" (college student) + ''0.003*"komunikasi" (communication)'	Publik masih banyak tertipu oleh judi online bahkan korban banyak dari mahasiswa (Online gambling still deceives the public; many victims are students.)
1	'0.008*"investasi" (investment) + 0.006*"berkedok" (under the guise) + 0.005*"berinvestasi" (invest) + 0.004*"sahabat" (best friend) + 0.004*"ojk" (OJK is an abbreviation of Otoritas Jasa Keuangan: Financial Services Authority) + 0.004*"meresahkan" (unsettling) + 0.003*"masyarakat" (public) + 0.003*"berantas" (eradicate) + 0.003*"berhati" (be careful) + 0.003*"phising"	Penipuan yang berkedok investasi meresahkan ojk mengharuskan masyarakat lebih berhati-hati sehingga tidak mengalami phising (Fraud under the guise of investment is troubling, and the OJK requires people to be more careful so they don't experience phishing)

Each model describes each topic. From the topic above, several words are the same but have different probability values. As in the topic, models 1 and 3 have the word "society" with different probability values. The order is based on the importance of each topic to the corpus, such as topic-6 which has a marginal topic of 22.7%, topic 2 of 19.1%, topic 5 of 16.3%, topic 4 of 13.5%, topic 3 of 11.6%, topic 7 of 9.4%, and topic 1 of 7.4%.

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the research that has been done, an analysis of tweet data related to the keyword "penipuan" on Twitter social media was carried out. The total tweet data collected is 11.581. After cleaning the data, the weighting stage uses the id2word—doc2bow method to produce a text representation of Term Frequency (TF). The analysis results show that the 7th model has the highest coherence value of 0.491899.

Furthermore, an analysis was carried out using the LDA (Latent Dirichlet Allocation) method. The research results show that there are some relationships between the identified topics. Topic 3 intersects with topics 7, 1, and 4. Topic 1 intersects with Topics 3 and 7, while Topic 7 intersects with Topics 1 and 3. That shows that related topics are related to one another.

Based on the results of the LDA analysis, topics can be arranged according to the level of importance of each topic to the corpus. Topic 6 has the highest level of marginal interest, namely 22.7%. Topic 2 was ranked second with an importance level of 19.1%, followed by Topic 5 with an importance level of 16.3%. Topic 4 has an importance level of 13.5%, while Topic 3 is 11.6%. Topic 7 has an interest rate of 9.4%, and Topic 1 has the lowest interest rate of 7.4%.

Therefore, the analysis using the LDA method has provided a more comprehensive understanding of fraudulent topics that often appear on the Twitter platform based on the analyzed tweet data.

Suggestion

Based on the in-depth analysis conducted on the results and discussion obtained from this research and the conclusions drawn, it is suggested to implement a more proactive strategy to obtain maximum results. One of the steps that can be taken is to add and select words from the stopword list. Identifying less relevant stopwords can help analyze the results obtained by studying the list of stopwords used more deeply and enriching the list with more specific words.

REFERENCES

Bhat, M. R., Kundroo, M. A., Tarray, T. A., & Agarwal, B. (2020). Deep Lda : A New Way To Topic Model. Journal Of Information And Optimization Sciences, 41(3), 823–834. Https://doi.Org/10.1080/02522667.2019.1616911

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An Evaluation Of Document Clustering And Topic Modelling In Two Online Social Networks: Twitter And Reddit. Information Processing And Management, 57(2). Https://doi.Org/10.1016/J.Ipm.2019.04.002

Fahlevvi, M. R., & Sn, A. (2022). Topic Modeling On Online News.Portal Using Latent Dirichlet Allocation (LDA). Ijccs (Indonesian Journal Of Computing And Cybernetics Systems), 16(4), 335. Https://doi.Org/10.22146/Ijccs.74383

Gupta, H., & Patel, M. (2021). Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert. Proceedings - International Conference On Artificial Intelligence And Smart Systems, Icais 2021, 511–517. Https://doi.Org/10.1109/Icais50930.2021.9395976

Hafis, F. (2020). Apa Yang Harus Dilakukan Jika Jadi Korban Penipuan Online? Ini Solusi Kominfo. Https://Www.Kominfo.Go.Id/Content/Detail/27912/Apa-Yang-Harus-Dilakukan-Jika-Jadi-Korban-Penipuan-Online-Ini-Solusi-Kominfo/0/Sorotan_Media

Kannitha, D. Z. T., Mustafid, & Kartikasari, P. (2022). Pemodelan Topik Pada Keluhan Pelanggan Menggunakan Algoritma Latent Dirichlet Allocation Dalam Media Sosial Twitter. 11(2), 266–277. Https://Ejournal3.Undip.Ac.Id/Index.Php/Gaussian/

Khatulistiwa, G. (2021). Heboh Penipuan Berkedok Donasi Di Twitter, Ini Tips Hindari Tindak Kejahatan Ini Di Internet. Https://Journal.Sociolla.Com/Lifestyle/Tips-Hindari-Penipuan-Internet

Kurnia, N., Rahayu, Wendratama, E., Monggilo, Z. M. Z., Damayanti, A., Angendari, D. A. D., Abisono, F. Q., Shafira, I., & Desmalinda. (2022). Penipuan Digital Di Indonesia Modus, Medium, Dan Rekomendasi.

Mandal, K. (2020). Topic Modeling: Techniques And Ai Models. Https://Dzone.Com/Articles/Topic-Modelling-Techniques-And-Ai-Models

Nur, E. (2021). Peran Media Massa Dalam Menghadapi Serbuan Media Online The Role Of Mass Media In Facing Online Media Attacks.

Nurlayli, A., & Nasichuddin, M. A. (2019). Topic Modeling Penelitian Dosen Jptei Uny Pada Google Scholar Menggunakan Latent Dirichlet Allocation. 4(2), 154–161. Https://doi.Org/10.21831/Elinvo.V4i2

Polyzos, E., & Wang, F. (2022). Twitter And Market Efficiency In Energy Markets: Evidence Using Lda Clustered Topic Extraction. Energy Economics, 114. Https://doi.Org/10.1016/J.Eneco.2022.106264

Kaila, R. P. (2020). Informational Flow On Twitter-Corona Virus Outbreak-Topic Modelling Approach. International Journal Of Advanced Research In Engineering And Technology (Ijaret), 11(3), 128–134. Http://Www.Iaeme.Com/Ijaret/Index.Asp128http://Www.Iaeme.Com/Ijaret/Issues.Asp?Jtype=Ijaret&Vtype=11&Itype=3journalimpactfactor

Puspitasari, I. (2018). Pertanggungjawaban Pidana Pelaku Tindak Pidana Penipuan Online Dalam Hukum Positif Di Indonesia Oleh. 8(Mei), 1–14. Https://Id.Wikipedia.Org/Wiki/Globalisasi

Sasmita, R. A., & Falani, A. Z. (2018). Pemanfaatan Algoritma Tf/Idf Pada Sistem Informasi Ecomplaint Handling. Jurnal Link, 27(1).

Saura, J. R., Reyes-Menendez, A., & Palos-Sanchez, P. (2019). Are Black Friday Deals Worth It? Mining Twitter Users' Sentiment And Behavior Response. Journal Of Open Innovation: Technology, Market, And Complexity, 5(3). Https://doi.Org/10.3390/Joitmc5030058

Shalihah, N. F. (2021). Ramai Penipuan "Cancel Order" Di Twitter, Bagaimana Menyiasatinya? Https://Www.Kompas.Com/Tren/Read/2021/02/17/123200865/Ramai-Penipuan-Cancel-Order-Di-Twitter-Bagaimana-Menyiasatinya-?Page=All

Silveira, R., Fernandes, C. G. O., Neto, J. A. M., Furtado, V., Ernesto, J., & Filho, P. (2021). Topic Modelling Of Legal Documents Via Legal-Bert 1.

Suhartono, D. (2018). Latent Dirichlet Allocation (LDA). Https://Socs.Binus.Ac.Id/2018/11/29/Latent-Dirichlet-Allocation-Lda/

Trinugraheni, N. F. (2022). Waspada Modus Penipuan Terbaru Di Twitter, Manfaatkan Kesuksesan Nft Moonbirds. Https://Www.Tribunnews.Com/Techno/202

2/04/19/Waspada-Modus-Penipuan-Terbaru-Di-Twitter-Manfaatkan-Kesuksesan-Nft-Moonbirds

Wahyudi, D., Sugiarto Samosir, H., & Sintha Devi, R. (2022). Akibat Hukum Bagi Pelaku Tindak Pidana Penipuan Online Melalui Modus Arisan Online Di Media Sosial Elektronik.

Yuwita, N., Mauhibatillah, N., & Ulyah, H. '. (2022). Dramaturgi: Budaya Flexing Berkedok Penipuan Di Media Sosial (Studi Kasus Indra Kenz Dan Doni Salmanan). Jurnal Komunikasi Dan Media, 7(1).

Zhou, Z., Qin, J., Xiang, X., Tan, Y., Liu, Q., & Xiong, N. N. (2020). News Text Topic Clustering Optimized Method Based On Tf-Idf Algorithm On Spark. Computers, Materials And Continua, 62(1), 217–231. Https://doi.Org/10.32604/Cmc.2020.06431

Zvornicanin, E. (2022, January 30). Topic Modeling And Latent Dirichlet Allocation (Lda). Https://Datascienceplus.Com/Topic-Modeling-And-Latent-Dirichlet-Allocation-Lda/