Text Mining Techniques for Identify Islamophobic Conversation Language by Selecting Preprocessing Feature

Fachrul Kurniawan (✉ fachrulk@ti.uin-malang.ac.id)
Universitas Islam Negeri Maulana Malik Ibrahim
Badruddin
Universitas Islam Negeri Maulana Malik Ibrahim
Aji Prasetya Wibawa
State University of Malang

Research Article

Keywords: social media, Islamophobia, language, text mining

Posted Date: December 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1105114/v1

Abstract

By identifying a text's polarity, sentiment analysis is a technique for extracting information from a person's attitude about an issue or occurrence. The grouping is made to discuss whether the reader is positive or negative. The drop duplication procedure creates 4339 from the preceding 10997, and the result language detection is 31 languages, thanks to the pre-processing stage. Although the data comes from the world's largest Muslim country, the problem is not limited to it, as evidenced by the use of text mining tools to identify languages.

Introduction

The usage of social media to have dialogues has expanded in recent years [1], [2]. This corresponds to the exponential growth of the internet, which continues to increase year after year. The concerns that arise in social media dialogues are diverse, ranging from positive to bad. Islamophobia is one of the essential concerns to consider. This problem develops as a result of religious conflict and the rapid advancement of digital technology [3]–[5].

Hatred of religion stems from violence committed solely to protect one's truth and the result of advances in communication technology, which allows all information to be disseminated fast and without verification. Extremely prejudiced words are thought to influence communication amongst fellow believers. As the world's largest Muslim country, Indonesia is keenly interested in Islamophobia, which can have political and security implications [6], [7].

The number of social media comments, particularly on Twitter. Positive and negative remarks are both acceptable. Islamophobia is one of the most often discussed harsh comments. Islamophobia is an outpouring of anti-Islamic sentiment manifested in various ways, including demonstrations, the passage of legislation forbidding the display of Islamic symbols, and the dissemination of unfavorable views on social media [8], [9].

The process of community communication through social media will be fascinating to conduct an in-depth study of the classification of conversation content and community patterns that lead to sentences about Islamophobia [10]. Sentiment analysis is a technique or method used to identify how expressed using text and how that sentiment can be classified as positive or negative. The results of the prototype system achieve high precision (75-95% depending on the data) in finding sentiments on web pages and news articles [11].

Data on social media conversations with hashtags related to hate speech that causes fear of specific ideas, categorized by a city where the conversation took place, will aid in understanding society's social model and character. Activities involving the interaction between cutting-edge internet-based technology and society's participation in the growth of numerous disciplines that have not been fully realized, particularly in the area of ethics, have had exceptional social consequences. This study will take a quantitative approach, which stresses in-depth data analysis.

According to the problem description above, we will describe pre-processing in text mining to parse the composition of the sorts of language that frequently arise to determine which communities from which locations typically debate Islamophobia in this article. This identification is critical as the initial step toward lowering the social attitude associated with hating speech community debates. It is intended that a community will emerge to participate in the forums identified to comprehend the genuine teachings of religion better and lessen the hate speech community on social media.

Related Work

Islamophobia has become a hot topic in various circles, posing a challenge to individuals all around the world. This has an impact on many Muslims, particularly those who are minorities in specific locations. Western intellectuals invented "Islamophobia" to denote anti-Islamic sentiment and prejudice [12]. Islamophobia is a shorthand term for fear or mockery of the Islamic religion, as well as the hostility of most or all Muslims [13].

When connected to the internet, social media is a tool that can be used to interact or communicate online [14]–[16]. Social media is frequently used to form social relationships with others and share personal activities or real-life experiences. One of the elements fueling the development of social media is its high mobility and ease of use. In real-time, the community of social media users provides data in a wide range of unstructured formats and languages, as well as thoughts and attitudes [17]. One of the topics in data mining known as text mining is the availability of vast and unstructured data. This strategy is well-organized.

Social media platforms are very diverse in type and type, so that it allows people to choose the community they want. One of the platforms is Twitter, which provides several facilities for its users to interpret, convey, and share posts of up to 280 characters, better known as tweets. The platform is accessible via mobile devices, instant messaging, and website interface generating 326 million monthly active users [18]. By linking hashtags, users can share any information very quickly when they want to search for information. The Twitter social network is included in the speedy category in terms of information exchange due to its easy use and high mobility [19], [20].

The pre-processing stage involves determining the data's quality before it is processed using specialized algorithms to be categorized, classed, or visualized as required. This stage is also critical because it determines the data quality to be used. Some of the processes are governed by rules that the researchers specify. Imperfect data, data interference, and inconsistent data can all be avoided by pre-processing. Pre-processing is critical in sentiment analysis, particularly in social media, where informal and unstructured words or sentences abound, as well as a lot of noise [21]–[23].

The following are the preprocessing stages often carried out, namely case folding, punctuation removal, tokenizing, and stop words removal. Case folding is a way to convert data in the same font size, which converts all into lowercase letters. punctuation removal, which is the stage to remove punctuation marks, numbers, links, and others. There are punctuation marks in some conversation data such as periods (.), commas (,), and a link. It is not necessary, so it needs to be removed. The process of dividing sentences into words and forming word vectors is a tokenizing process. Elimination of irrelevant words reduces the repetition of words that occur to give rise to unconnected opinions [24].

The Purpose Methods

The overall scheme research activity in this study includes five pre-processing features and a selection and visualization detecting language. Figure 1 presents the scheme activity of this research.

Data Crawling

In this study, the data came from social media, which became a place of reference for community conversations and especially religious or belief communities. The data comes from retrieval that comes from the Islamophobic hashtag on a social media platform called Twitter. By using the crawling technique carried out through the language

Selection and Feature Preprocessing

The preprocessing selection process starts from identifying or determining features by considering the interests of the desired research target. The selection uses five stages, namely drop duplicate, cleaning, case folding, tokenizing, stemming. The determination of the features used is the researcher's choice according to the level of importance of the data sorting process.

Language Detection

This study resulted in detecting the number of languages that often appear and are used in conversations which detected the number of languages that are most widely used. This language detection will use a python-based algorithm, which produces several frequently used languages and visualizations in graphical form.

Experiment

The data used in this study are social media users' opinions with the hashtag Islamophobia in 2021. The data taken is only in the form of tweets in all languages. Furthermore, the following is the dataset that has been collected, shown in Table 1.

Table 1
Dataset crawling Islamophobia in Bahasa Indonesia

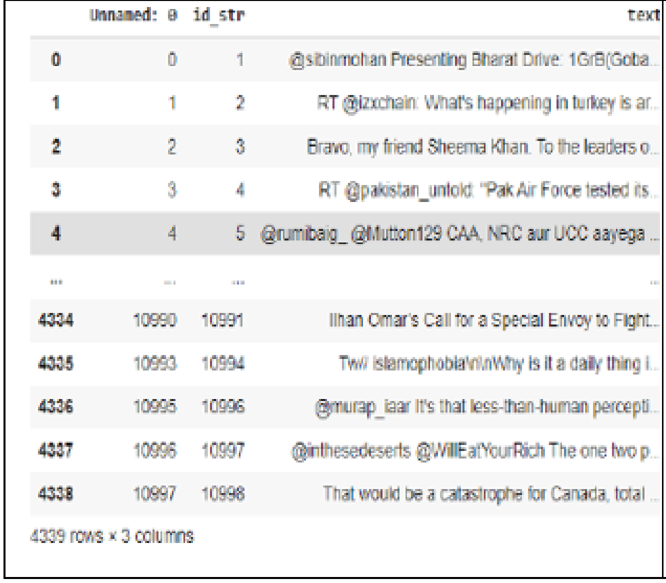
No.	Tweet
1	rintangan terbesar tangani islamofobia adalah perbedaan definisi dan konteks https://t.co/5c4chyv9zd
2	islamofobia di Eropa diduga disuburkan oleh golongan elite ini. https://t.co/h5yquetakc
3	beberapa muslim inggris mengalami rasialisme, islamofobia, dan pengucilan. https://t.co/7id5wncsge
4	islam agama masa depan, agama rahmat bagi semesta alam dan agama sejalan dgn peradaban dunia yg manusiawi, muhammad saw adalah role model /reladan karena salah kaprah itu ia pun bersemangat untuk mengampanyekan islamofobia melalui akunakun media sosial - https://t.co/bzetbkxerc

The experimental process was carried out, starting with dropping duplicates, namely by deleting data detected as word duplication. The data used is 10997 tweet data, and after it is done, drop duplicate data to 4339 data. Then do the deletion of duplicate data first. The data used is 10,997 tweet data, and after dropping duplicates, it becomes 4339 data. After the data has been determined, the next step is the language detection process. This process is carried out for labeling and calculating the number of languages used in tweets. This process found that the most widely used language was English with 3854 tweets, followed by Indonesian with 143 tweets and 31 tweets from other languages.

The cleaning process carried out in the preprocessing is deleting emoticons/characters, URLs, mentions, hashtags, numbers, excess spaces, retweets, punctuation marks, and enter. It is the process of deleting symbols and characters contained in the data used in this study. The next step is tokenizing, which is the process of making tweet sentences per word. The last process is stemming, which is removing affixes, both prefixes, infixes, and suffixes. Moreover, the following is shown in Table 2 and 3, the selection process for the preprocessing feature.

Table 2. The selection process features for Cleaning, Case Folding, and Tokenizing

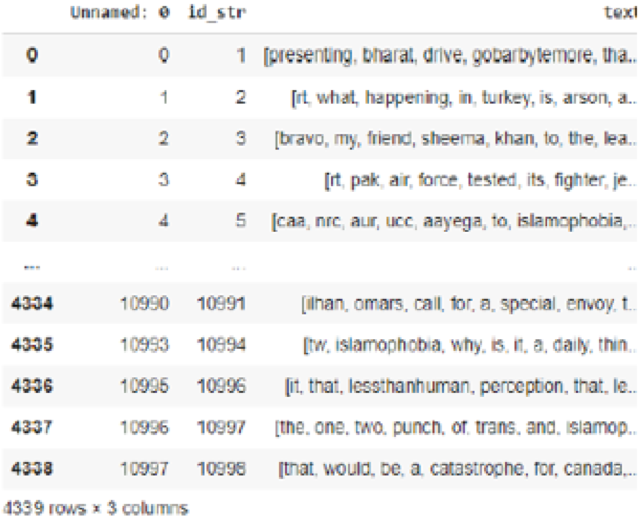
Before



	Unnamed: 0	id_str	text
0	0	1	@sibinmohan Presenting Bharat Drive: 1GrB(Goba...
1	1	2	RT @izxchain: What's happening in turkey is ar...
2	2	3	Bravo, my friend Sheema Khan. To the leaders o...
3	3	4	RT @pakistan_untold: "Pak Air Force tested its...
4	4	5	@rumibaig_ @Mutton129 CAA, NRC aur UCC aayega ...
...	...	...	...
4334	10990	10991	Ilhan Omar's Call for a Special Envoy to Fight...
4335	10993	10994	Tw/ islamophobia\n\nWhy is it a daily thing i...
4336	10995	10996	@murap_iaar It's that less-than-human percepti...
4337	10996	10997	@inthesedeserts @WillEatYourRich The one two p...
4338	10997	10998	That would be a catastrophe for Canada, total ...

4339 rows × 3 columns

After

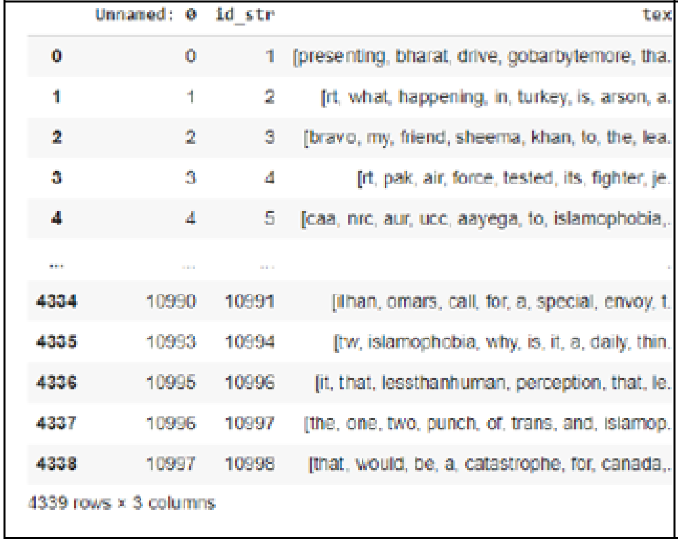


	Unnamed: 0	id_str	text
0	0	1	[presenting, bharat, drive, gobarbytemore, tha...
1	1	2	[rt, what, happening, in, turkey, is, arson, a...
2	2	3	[bravo, my, friend, sheema, khan, to, the, lea...
3	3	4	[rt, pak, air, force, tested, its, fighter, je...
4	4	5	[caa, nrc, aur, ucc, aayega, to, islamophobia,...
...	...	...	...
4334	10990	10991	[ilhan, omars, call, for, a, special, envoy, t...
4335	10993	10994	[tw, islamophobia, why, is, it, a, daily, thin...
4336	10995	10996	[it, that, lessthanhuman, perception, that, le...
4337	10996	10997	[the, one, two, punch, of, trans, and, islamop...
4338	10997	10998	[that, would, be, a, catastrophe, for, canada,...

4339 rows × 3 columns

Table 3. The selection process for stemming

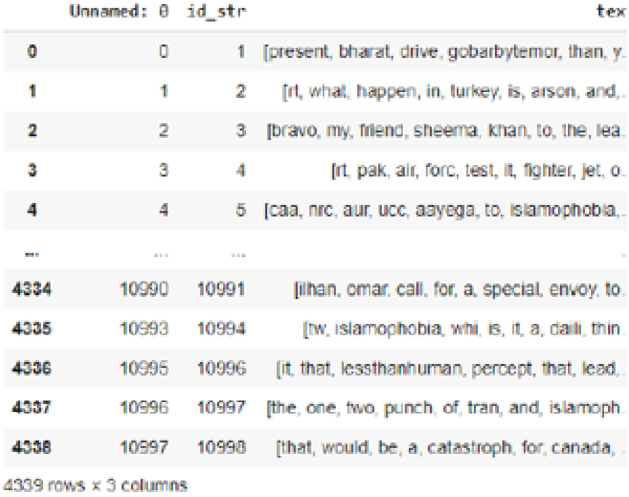
Before



	Unnamed: 0	id_str	tex
0	0	1	[presenting, bharat, drive, gobarbytemore, tha.
1	1	2	[rt, what, happening, in, turkey, is, arson, a.
2	2	3	[bravo, my, friend, sheema, khan, to, the, lea.
3	3	4	[rt, pak, air, force, tested, its, fighter, je.
4	4	5	[caa, nrc, aur, ucc, aayega, to, islamophobia,.
...	...	...	.
4334	10990	10991	[ilhan, omars, call, for, a, special, envoy, t.
4335	10993	10994	[tw, islamophobia, why, is, it, a, daily, thin.
4336	10995	10996	[it, that, lessthanhuman, perception, that, le.
4337	10996	10997	[the, one, two, punch, of, trans, and, islamop.
4338	10997	10998	[that, would, be, a, catastrophe, for, canada,.

4339 rows × 3 columns

After



	Unnamed: 0	id_str	tex
0	0	1	[present, bharat, drive, gobarbytemor, than, y.
1	1	2	[rt, what, happen, in, turkey, is, arson, and,.
2	2	3	[bravo, my, friend, sheema, khan, to, the, lea.
3	3	4	[rt, pak, air, forc, test, it, fighter, jet, o.
4	4	5	[caa, nrc, aur, ucc, aayega, to, islamophobia,.
...	...	...	.
4334	10990	10991	[ilhan, omar, call, for, a, special, envoy, to.
4335	10993	10994	[tw, islamophobia, whi, is, it, a, daili, thin.
4336	10995	10996	[it, that, lessthanhuman, percept, that, lead,.
4337	10996	10997	[the, one, two, punch, of, tran, and, islamoph.
4338	10997	10998	[that, would, be, a, catastroph, for, canada, .

4339 rows × 3 columns

The selection process of the preprocessing feature shows a very significant result where all the data, which initially amounted to 10997, turned into 4339 data. Feature selection is selected and adjusted to make the data better and more accurate.

Discussion

From the experimental results, some observations about the proposed approach are shown as follows. First, selecting the preprocessing feature can be carried out effectively where the time required to determine the preprocessing technique can be. Second, data taken from specific social media requires the selection of preprocessing according to interests. The selection of preprocessing features improves text mining-based data structuring performance to become more ready to be processed in any form. However, the performance of feature selection needs to be compared between each stage to avoid inefficiency in the process of each stage. In the end, crawling data originating from conversation forums on social media requires processing using a selection of preprocessing features because each platform requires different treatment. The behavior outlined in the discussion of Islamophobia by social media users does not necessarily come from the majority of countries that use religion as one of the parameters in every community activity.

Conclusion

This research shows that social media, especially Twitter, has become a commonly used platform for conversations about the issue of Islamophobia. This paper proposes selecting the preprocessing feature to control the gross crawling data and then measure it in a limited and short manner. The selection of features used in crawling data processing can optimize results that can be used to perform limited analysis. Research using preprocessing features with the hashtag Islamophobia identified 31 types of languages from various countries. Identification uses text mining techniques with very maximum feature selection to find out the target desired by the researcher in a limited and fast manner. In the future, additional features will be produced in pre-processing in text mining to create high-quality data.

Declarations

Acknowledgements

This work was supported in part by the research institute and community service LP2M Universitas Islam Maulana Malik Ibrahim Malang 2021.

References

1. Eriksson, M. & “ Lessons for Crisis Communication on Social Media: A Systematic Review of What Research Tells the Practice., ” *Int. J. Strateg. Commun*, **12** (no. 5), 526–551 https://doi.org/doi: 10.1080/1553118X.2018.1510405. (Oct. 2018).
2. Wang, Y., Yang, Y. & “ Dialogic communication on social media: How organizations use Twitter to build dialogic relationships with their publics., ” *Comput. Human Behav*, **104**, 106183 https://doi.org/doi: 10.1016/j.chb.2019.106183. (Mar. 2020).
3. Eckert, S., Metzger-Riftkin, J., Kolhoff, S., O’Shay-Wallace, S. & “ A hyper differential counterpublic: Muslim social media users and Islamophobia during the 2016 US presidential election., ” *New Media Soc*, **23** (no. 1), 78–98 https://doi.org/doi: 10.1177/1461444819892283. (Jan. 2021).

4. Vidgen, B., Yasseri, T. & “ Detecting weak and strong Islamophobic hate speech on social media., ” *J. Inf. Technol. Polit*, **17** (no. 1), 66–78 https://doi.org/doi: 10.1080/19331681.2019.1702607. (Jan. 2020).
5. Hashmi, U. M., Rashid, R. A. & Ahmad, M. K. “The representation of Islam within social media: a systematic review,” *Information, Commun. Soc.*, vol. 24, no. 13, pp. 1962–1981, Oct. 2021, doi: 10.1080/1369118X.2020.1847165
6. Shukri, S. F. M. & “ The Perception of Indonesian Youths toward Islamophobia: An Exploratory Study., ” *Islam. Stud. J*, **5** (no. 1), 61 https://doi.org/doi: 10.13169/islastudj.5.1.0061. (2019).
7. Syarif, Z., Mughni, S. A. & Hannan, A. “Post-truth and Islamophobia narration in the contemporary Indonesian political constellation,” *Indones. J. Islam Muslim Soc.*, vol. 10, no. 2, pp. 199–225, Dec. 2020, doi: 10.18326/ijims.v10i2.199-225
8. Allen, C. *Islamophobia* (Routledge, London, 2016).
9. Sayyid, S. & “ A Measure of Islamophobia., ” *Islam. Stud. J*, **2** (no. 1, p. 10,), https://doi.org/doi: 10.13169/islastudj.2.1.0010. (2014).
10. Rani, S., Kumar, P. & “ Deep Learning Based Sentiment Analysis Using Convolution Neural Network., ” *Arab. J. Sci. Eng*, **44** (no. 4), 3305–3314 https://doi.org/doi: 10.1007/s13369-018-3500-z. (Apr. 2019).
11. Nasukawa, T. & Yi, J. “Sentiment analysis,” in *Proceedings of the international conference on Knowledge capture - K-CAP ’03*, 2003, p. 70, doi: 10.1145/945645.945658
12. Sirgy, M. J., Kim, M. Y., Joshanloo, M. & Bosnjak, M. “Is Subjective Ill-Being Related to Islamophobia in Germany? In Search for Moderators,” *J. Happiness Stud.*, vol. 20, no. 8, pp. 2655–2675, Dec. 2019, doi: 10.1007/s10902-018-0063-3
13. Roose, J. M. & Turner, B. S. “Islamophobia, Science and the Advocacy Concept,” *Society*, vol. 56, no. 3, pp. 210–221, Jun. 2019, doi: 10.1007/s12115-019-00357-6
14. Sendari, S., Zaeni, I. A. E., Lestari, D. C. & Hariyadi, H. P. “Opinion Analysis for Emotional Classification on Emoji Tweets using the Naïve Bayes Algorithm,” *Knowl. Eng. Data Sci.*, vol. 3, no. 1, pp. 50–59, Aug. 2020, doi: 10.17977/um018v3i12020p50-59
15. Huang, Y. T., Su, S. F. & “ Motives for Instagram Use and Topics of Interest among Young Adults., ” *Futur. Internet*, **10** (no. 8), 77 https://doi.org/doi: 10.3390/fi10080077. (2018).
16. Baruah, T. D. & “ Effectiveness of Social Media as a tool of communication and its potential for technology enabled connections: A micro-level study., ” *Int. J. Sci. Res. Publ*, **2** (no. 1), 1–10 (2012)., doi: ISSN 2250–3153
17. Hitesh, M., Vaibhav, V., Kalki, Y. A., Kamtam, S. H. & Kumari, S. “Real-Time Sentiment Analysis of 2019 Election Tweets using Word2vec and Random Forest Model,” in 2019 *2nd International Conference on Intelligent Communication and Computational* Techniques *(ICCT)*, Sep. 2019, pp. 146–151, doi: 10.1109/ICCT46177.2019.8969049
18. Saad, S. E., Yang, J. & “ Twitter Sentiment Analysis Based on Ordinal Regression., ” *IEEE Access*, **7**, 163677–163685 https://doi.org/doi: 10.1109/ACCESS.2019.2952127. (2019).

19. Jianqiang, Z., Xiaolin, G., Xuejun, Z. & " Deep Convolution Neural Networks for Twitter Sentiment Analysis., " *IEEE Access*, **6**, 23253–23260 https://doi.org/doi: 10.1109/ACCESS.2017.2776930. (2018).

20. Naseem, U., Razzak, I., Musial, K. & Imran, M. "Transformer based Deep Intelligent Contextual Embedding for Twitter sentiment analysis," *Futur. Gener. Comput. Syst.*, vol. 113, pp. 58–69, Dec. 2020, doi: 10.1016/j.future.2020.06.050

21. Camacho-Collados, J. & Pilehvar, M. T. "On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis," Jul. 2017, [Online]. Available: http://arxiv.org/abs/1707.01780

22. Pradha, S., Halgamuge, M. N. & Vinh, N. T. Q. "Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data," in 2019 *11th International Conference on Knowledge and Systems Engineering (KSE)*, Oct. 2019, pp. 1–8, doi: 10.1109/KSE.2019.8919368

23. Alam, S. & Yao, N. "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Comput. Math. Organ. Theory*, vol. 25, no. 3, pp. 319–335, Sep. 2019, doi: 10.1007/s10588-018-9266-8

24. Moolthaisong, K. & Songpan, W. "Emotion Analysis and Classification of Movie Reviews Using Data Mining," in *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, Jul. 2020, pp. 89–92, doi: 10.1109/DATABIA50434.2020.9190363

Figures

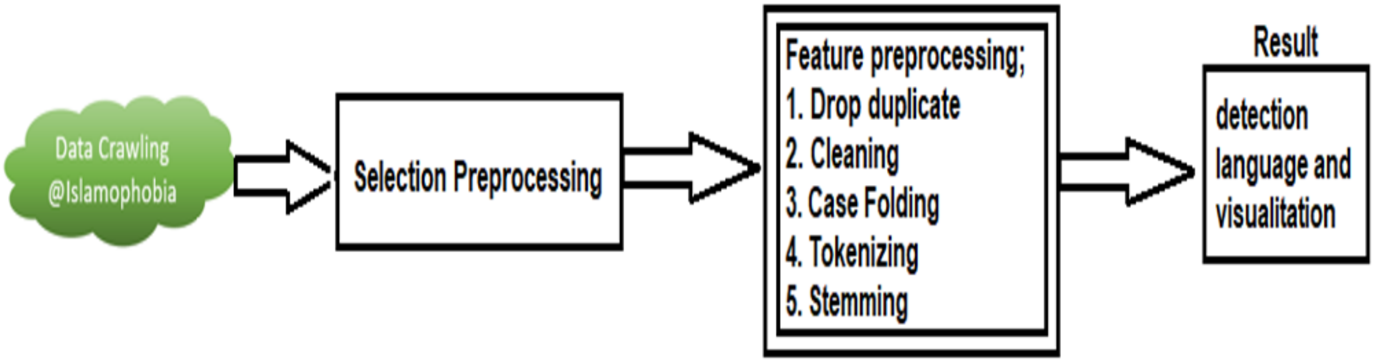


Figure 1

The scheme activity of the proposed research