

Spam Detection Using Word Sense Disambiguation

Ms. Anjana B Sanjada¹ Dr. Kinjal Adhvaryu² Ms. Manisha Chaudhari³ Ms. Rama M. Maliya⁴
Ms. Pooja Shah⁵

^{1,2,5}Shankersinh Vaghela Bapu Institute of Technology, India ³Government Polytechnic Gandhinagar, India ⁴Sharda University, Greater Noida, India

Abstract— with advancement in technology most of the fashionable day communication takes place through emails. This has made the method of communication much faster and easier because it saves time. One probable disadvantage of using emails as a major mode communication is advertisements. Many companies use them for his or her advertisement and keep it up sending emails that contain unwanted information often mentioned as Spam. Although many approaches are developed for the identification of spam emails but none of them gives good accuracy in spam identification and screening. In existing work two semantic level analysis is used. And in this proposed work Word Sense Disambiguation (WSD) used for spam detection by using CN2 SD algorithm.

Keywords: Word Sense Disambiguation (WSD)

I. INTRODUCTION

With the fast and easy use of internet, electronic mails have end up to be an efficient, rapid and cheap means of communication. By using emails, a user can transfer his/her data anywhere within the world with in seconds. The rise in usage of emails over the past few years has led to a subsequent increase within the spam emails. Spam is typically defined as unwanted or unsolicited mails that are send bulk to different recipients and network of virus infected computers. Word-sense disambiguation (WSD) is a concerned with identifying which sense of a word is used in a sentence. Word sense disambiguation is the process of selecting the most correct sense of the word in given sentence. The use of word in a particular context. Lexical ambiguity, semantics etc. When we have bulk of all mails, it is not possible to find out spam & ham mails from it.

Semantic Analyzer is one of the phase of Natural Language Processing, and this analyser checks about the word is correct or not. As we know in normal text same word having different meaning. This type of ambiguity is one the biggest problem and this type of semantic ambiguity also known as WSD- Word Sense Disambiguation. This research paper is all about find out Spam & Ham Email from given dataset using WSD approach.

II. RELATED WORK

- 1) A semantic-based classification approach for an enhanced spam detection: In this paper focus a method based on two semantic level analysis. First email categorization by domain, second email semantic features extraction. we investigate the utilize of a content semantic examination to move forward the precision of spam detection. We propose a strategy based on two semantic level examination. Within the to begin with level, we categorize emails by specific spaces (e.g., Wellbeing, Instruction, Back, etc.) to empower a partitioned conceptual see for spams in each space.

Within the moment level, we combine a set of manually-specified and automatically-extracted semantic highlights for spam discovery in each space. These highlights are implied to summarize the email content into compact subjects separating spam from non-spam emails in an effective way.

- 2) Opinion spam detection: Using multi-iterative graph-based model: In this paper focus a novel graph based model called MGSD by increasing the use of online reviews relating to purchasing decision of consumer analysis market trends.
- 3) Spam filtering using a logistic regression model trained by an artificial bee colony algorithm: In this paper focus a novel spam detection method that combine the artificial bee colony algorithm with a logistic regression classification model.

III. PROPOSED WORK

A. WSD Approach:

Word Sense Disambiguation (WSD) is an important but challenging technique in the area of natural language processing (NLP). Hundreds of WSD algorithms and systems are available, but less work has been done in regard to choosing the optimal WSD algorithms. Various supervised learning, unsupervised learning and semi-supervised techniques have been discussed. WSD is mainly used in Information Retrieval (IR), Information Extraction (IE), Machine Translation (MT), Question Answering (QT), Content Analysis, Word Processing, Lexicography and Semantic Web.

- 1) WSD APPROACHES: There are two approaches that are followed for Word Sense Disambiguation (WSD): Machine-Learning Based approach and Knowledge Based approach. In Machine learning based approach, systems are trained to perform the task of word sense disambiguation. In Knowledge based approach, it requires external lexical resources like Word Net, dictionary, thesaurus etc.

B. Dictionary Based Algorithm:

They do not use corpora evidences for disambiguation. The Lesk method is the seminal dictionary-based method introduced by Michael Lesk in 1986. The Lesk definition, on which the Lesk algorithm is based is “measure overlap between sense definitions for all words in context”. However, in 2000, Kilgariff and Rosensweig gave the simplified Lesk definition as “measure overlap between sense definitions of word and current context”, which further means identify the correct sense for one word at a time. Here the current context is the set of words in surrounding sentence or paragraph.

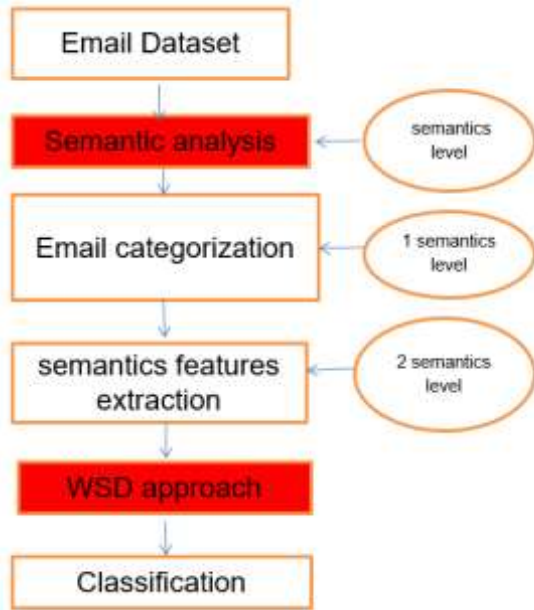


Fig. 1: system flow

C. Applications of Word Sense Disambiguation (WSD)

- 1) Machine Translation
- 2) Information Retrieval (IR)
- 3) Text Mining and Information Extraction (IE)
- 4) Lexicography

Steps:

- 1) Read Dataset
- 2) Preprocessing
- 3) Training
- 4) Testing

IV. RESULTS

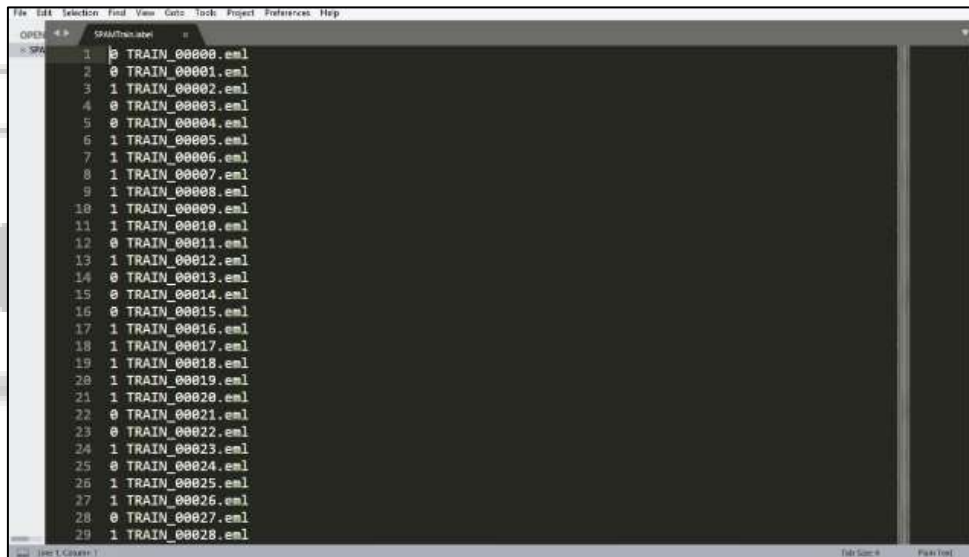


Fig. 2: Spam Train

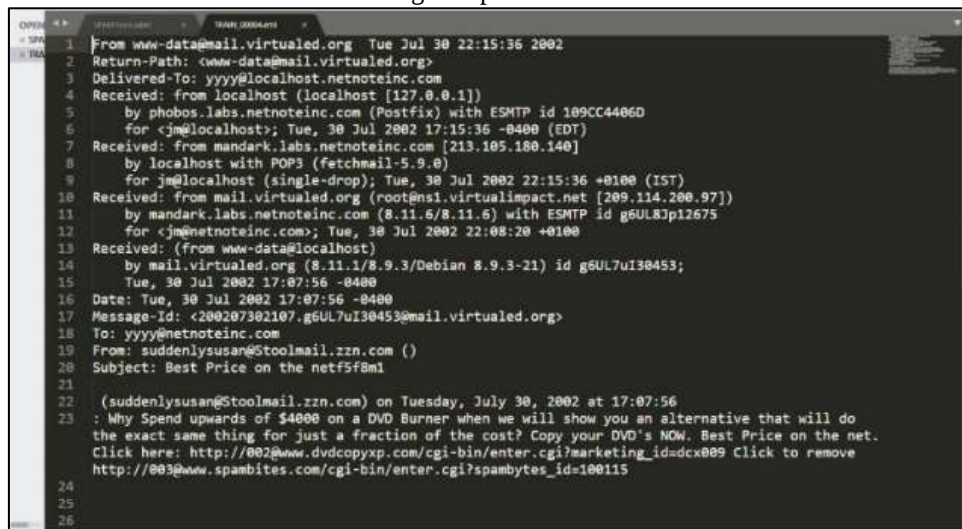


Fig. 3: eml file



Fig. 4: dictionary

V. CONCLUSION

In early stage of proposed method Anomaly detection used for filtering and enables a better spam detection compared to existing methods and leads to more interpretable results.

REFERENCES

- [1] A semantic-based classification approach for an enhanced spam detection Computers & Security Elsevier Ltd. January-2020 Nadjate Saidani *, Kamel Adi , Mohand Saïd Allili.
- [2] Opinion spam detection: Using multi-iterative graph-based model Information Processing and Management Elsevier -2020 Shirin Noekhaha,b,*, Naomie binti Salima, Nor Hawaniah Zakaria.
- [3] Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. Applied Soft Computing Journal, Elsevier March 2020 Author: Bilge Kagan Dedeturk a,*, Bahriye Akay b.
- [4] SimilCatch: Enhanced social spammers detection on Twitter using Markov Random Fields-2020 Information Processing and Management Elsevier 2020 Author: Nour El-Mawass*,a, Paul Honeinea, Laurent Vercouterb.
- [5] Detection of spam and threads identification in E-mail spam corpus using content based text analytics method-April 2020 Materials Today: Proceedings Elsevier April 2020 Author: U. Murugavel , R. Santhi.
- [6] SPAMI: A cognitive spam protector for advertisement malicious Images-May 2020 Information Sciences, Elsevier June 2020 Author: Aisha Makkar a, Neeraj Kumar b,c,d,†, Albert Y Zomaya e, Shalini Dhiman.
- [7] An overview of online fake news: Characterization, detection, and discussion Information Processing and Management ELSEVIER Author: Xichen Zhang*, Ali A. Ghorbani Ca.
- [8] An efficient deep learning-based scheme for web spam detection in IoT environment Future

Generation Computer Systems Elsevier-March 2020 Author: Aisha Makkar a, Neeraj Kumar

- [9] A deep learning model for Twitter spam detection Online Social Networks and Media Elsevier April 2020 Author: Zulfikar Alom a , Barbara Carminati b , *, Elena Ferrari .
- [10] Email spam detection using machine learning algorithm. IEEE Xplore Part Number: CFP20N67-RT; ISBN: 978-1-7281-5374-2 Sept 2020 Author: Nikhil Kumar.