International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 4, Issue 2, February 2015 **ISSN 2319 - 4847**

Direct and Indirect Discrimination Prevention in Data Mining By Using Natural Language Method

[1.]Ms Shraddha S Kediya, [2.]Prof S.V.Dabhade

[1.]Dept of Computer Sci
Shrimati Kashibai Navale College of Engg ,vadgaon,pune

[2.]Dept of Computer Sci
Shrimati Kashibai Navale College of Engg ,Vadgaon,Pune

ABSTRACT

In data mining, discrimination is a very important issue when considering the legal and ethical aspects of privacy preservation. It is more clear that most of the people do not have a wish to discriminated based on their race, nationality, religion, age and so on. This problem mainly arises when these kind of attributes are used for decision making purpose such as giving them a job and loan. Automatic data collection has become the most wanted method in the banking sector to make automatic decisions like loan granting/denial. Part-of-Speech (POS) tagging is the process of assigning a part-of-speech like noun, verb, adjective, adverb, or other lexical class marker to each word in a sentence. This paper presents a POS Tagger for English language text using Rule based approach, which will assign part of speech to the words in a sentence given as an input. n this paper we are going to use Natural Language Processing Approach for direct and indirect discrimination prevention. It consists of POS tagging and chunking methods. POS tagging is useful for identifying verbs, nouns, adjectives in a given line. On the basis of that we can identify the action words which may cause direct or indirect discrimination.

Keyword:- Direct discrimination, Indirect discrimination ,NLP,POS etc.

1.INTRODUCTION

Discrimination is termed as the act of unequally treating people on the basis of their belonging to a specific group. For example individuals may be discriminated because of their gender, ethnicity or nationality… etc. Different decision making tasks leads to the discrimination.eg loan granting/denial in the banking application .Discrimination is classified into two types. They are direct and indirect discrimination. Direct discrimination occurs when decisions are made based on the sensitive attributes. Indirect discrimination occurs when decisions are made based on the non-sensitive attributes.

Part-of-Speech (POS) tagging is the process of assigning a part-of-speech like noun, verb, adjective, adverb, or other lexical class marker to each word in a sentence. POS tagging is a necessary pre-module to other natural language processing tasks like natural language parsing, semantic analyzer, information extraction and information retrieval. A word can occur with different lexical class tags in different contexts. The main challenge in POS tagging involves resolving this ambiguity in possible POS tags for a word. We developed a POS tagger which will assign part of speech to the word in a sentence provided as input to the system. Here we have assigned five tags only viz. noun, adverb, adjective, verb and pronoun. Several approaches have been proposed and successfully implemented for POS tagging for different languages.

There are various approaches of POS tagging, which can be divided into three categories; rule based tagging, statistical tagging and hybrid tagging.

A. Rule based approach:

The rule based POS tagging model requires a set of hand written rules and uses contextual information to assign POS tags to words. The main drawback of rule based system is that it fails when the text is unknown, because the unknown word would not be present in the WordNet. Therefore the rule based system cannot predict the appropriate tags. Hence for achieving higher accuracy in this system we need to have an exhaustive set of hand coded rules.

B. Statistical approach:

A statistical approach includes frequency and probability. The simplest statistical approach finds out the most frequently used tag for a specific word from the annotated training data and uses this information to tag that word in the un annotated text. These systems are having more efficiency than the rule based approach. The problem with this approach is that it can come up with sequences of tags for sentences that are not acceptable according to the grammar rules of a language.

C. Hybrid approach:
A hybrid approach may perform better than statistical or rule based approaches. The POS tagger which is implemented using hybrid approach is having higher accuracy than the individual rule based or statistical approach. The hybrid approach first uses the set of hand coded language rules and then applies the probabilistic features of the statistical method. Most common POS taggers use a POS dictionary, which is also known as WordNet, having words tagged with a small set of possible output tags. In this paper we are study presenting the POS Tagger for Marathi Language .The main problem in part of speech tagging is ambiguous words. The Marathi Language is full of ambiguous words. There may be many words which can have more than one tag. To solve this problem we consider the context instead of taking single word.

2.LITERATURE SURVEY

J.Domingo-Ferrer et al. (2011) have developed a paper for rule protection for the indirect discrimination prevention in data mining. The datasets are trained and developed to make the classification rules to be extracted. Indirect discrimination rules cannot be extracted from the trained dataset. (i.e.) the trained datasets are free from indirect discrimination. Datasets are modified if any indirect discrimination occurs. Standard data mining algorithms are used to prevent the indirect discrimination from the training dataset.
Mykola Pechenizkiy et al. (2010) have developed a paper for discrimination aware decision tree learning. The decision tree models leads to the lower discrimination than the other models but with a little loss in the accuracy. The decision tree models are effective at removing the discrimination from the original datasets. The problem is the datasets are cleaned away for discrimination before the discovery of the classifier in the dataset.
Sara Hajian et al. (2011) have developed a paper for prevention of discrimination in data mining for intrusion and crime detection. Data mining algorithm are used to prevent the direct and indirect discrimination. The data set obtained is free from the discrimination. In addition to detect the discrimination intrusion fraud and crime is also detected in the given dataset.

3.RELATED WORK

EXISTING SYSTEM
The existing system is effective at removing the direct and indirect discrimination in the original dataset and preserves the data quality. The existing system does not require the standard data mining algorithm. They generally based on the classification rules of inductive part and reasoning on them the deductive part on the basis of the discriminative measures. Discrimination prevention methods are used in terms of the data quality and discrimination removal methods are used for both direct and indirect discrimination.
Drawbacks of the existing system are
(1) it takes more time to handle the decision tree. (2) It will not handle more data and cannot predict attributes (3) Mining data's are not trustworthy and system cannot handle more amount of data
PROPOSED SYSTEM
The proposed method can handle more data and discriminate them with the help of rule protection and generalization method. Preprocessing approach is used here. Different possible methods are compared for both direct and indirect discrimination method. Anti-discrimination methodology is introduced. Different measures of discriminating power of the mined decision rules are defined by the anti-discrimination. The unwanted memory space and the buffering memory are reduced. Discrimination free data models can be produced from the transformed dataset without seriously damaging the data quality.
In Proposed system more data can be discriminated. Discrimination is main thing of this process by the way of this process more people can serve without any partiality. Nondiscriminatory constraint is embedded into a decision tree learner by changing its splitting criterion
PROPOSED METHDOLOGY
The implementation part is mainly explained to prevent the discrimination in the dataset and to maintain the data quality for the given dataset.
A. Data Analysis:
Data analysis is to gather the data from the external disk. Dataset contains real life dataset and synthetic dataset. First of all we have to check if all the attributes are placed in a correct manner if any null values are present then those dataset attributes cannot be processed by the metric and other computation process. Data analysis is generally termed as the process of gathering and analysis of dataset individually in a given two dataset.
B. Utility Measures:
Utility measures are taken to remove the discrimination on the given dataset. Dataset are analyzed with certain measures to remove the discrimination from the specified data's. Indirect discrimination removal and measuring data quality of that process are computed by the mathematical functionality like metric and rule protection and rule generalization. With the use of these techniques and algorithm records are filter in short time.

D. Modifying Discriminatory Methods:

In this modified data's are converted into anonymized data. Anonymized data's means it does not behave like sensitive attributes but this data's can be processed. In this module modification are done on the sensitive attributes like gender, race, religion, sex, marital status and so on to anonymized the data's. In this module administrator can make the data to free from the sensitive attributes.

E. Decision Making:

Decisions could be depend on the attributes like gender, race and religion and so on. Each user gets score for their personal attributes in direct discrimination. Indirect discrimination can be done with the help of anonymized data's. Decisions should be done with the help of algorithm. The resulting data's are free from the discrimination removal and the data quality is maintained.

Discrimination prevention can be done by using three approaches

A) Pre-processing:- In this approach we used hierarchical based generalization, and transform the data in such a way that biased data are removed so that no unfair decision rule can be mined from the transformed data.

B) In-Processing: - In this approach the actual data mining algorithms is changes in such a way that the resulting models do not contain unfair decision rules. For example, an alternative approach to cleaning the discrimination from the original data set is proposed in whereby the nondiscriminatory constraint is embedded into a decision tree learner by changing its splitting criterion and pruning strategy through a novel leaf relabeling approach.

C) Post-processing:- In this approach the resulting data mining models is modify, instead of cleaning the original data set or changing the data mining algorithms. For example, in, a confidence-altering approach is proposed for classification rules inferred by CPAR algorithm.

Databases

1) WordNet:

WordNet is an electronic database which contains parts of speech of all the words which are stored in it. It is trained from the corpus for higher performance and efficiency.

2) **Corpus***:*

For correct POS tagging, training the tagger well is very important, which requires the use of well annotated corpora. Annotation of corpora can be done at various levels which include POS, phrase or clause level, dependency level etc. For POS Tagging in Marathi we are using a corpus which is based on tourism domain. It is an annotated corpus. As not much work done on Marathi language, we had to start with the unannotated corpus we took a small part of it and manually tag it.

3) Tagset

Apart from corpora, a well-chosen tagset is also important.

SYSTEM ARCHITECTUER

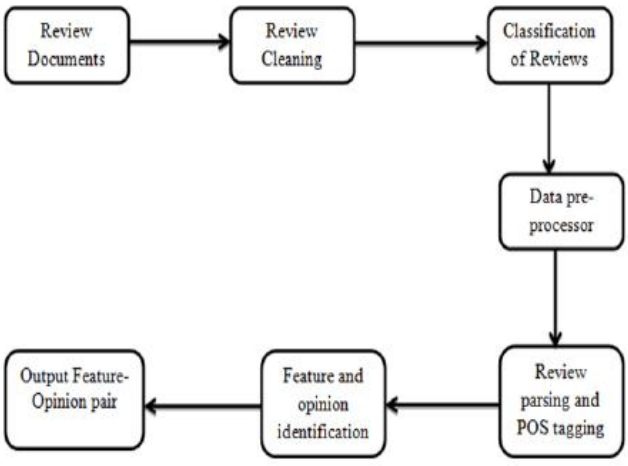


Fig 1 System Architecture

The need to identify and interpret possible difference in the linguistic style of text, such as formal or informal is increasing, as more and more people are using the Internet as their main research resource. In this section, the architecture and functional detail of the proposed opinion mining system to identify feature-opinion pairs is presented.

The proposed system takes care of the informal reviews. The idea of formal and informal words has been revised little in the proposed system. The use of English language can be grammatically incorrect with chances of spelling mistakes and wide use of shortcuts. Technically they can be called as formal opinions and informal opinions, where formal opinions refer to the use of proper English with no grammatical mistakes and informal refers to the improper use of English with grammatical mistakes, involving use of words which are not standard spellings e.g. "4get" instead of "forget". They say that the UK English is said to be formal English and US English is considered to be informal

English. Most of the customers use US English, hence it is very important to concentrate on these opinions too, or else they might be discarded as noise. Hence the system proposed deals with this aspect of informal reviews.

Table 1 List of Formal and Informal words

Formal	Informal
approximately	About
in addition	And
anyone	Anybody
request	ask for
employer	Boss
however	But
purchase	Buy
finish	End
sufficient	Enough
must	have to
are not	aren't
cannot	can't

Nowadays, most of the people have started using the sms language. Hence there is a large set of review data which will contain opinions in the short-form/sms language. Let us look at the following table to get an idea.

Table 2. Proposed list of formal and informal words

Formal	Informal
I am	M
Behave	Bhv/behv
Have	Hv
This	Dis
Good	Gud
Awesome	Awsum
Different	Diff
Bright	Brite
Working	Wrkin

In English POS tagger we use English WordNet as a tagset which will be working as our database. The
record in the tagset consists of two parts, first is the
word along with its intended tag and second is the root word for the corresponding word. The tag representation consists of 4 bits which represents
Noun, Adjective, Adverb, and Verb.

- When the first bit is 1 i.e. 1000 the word is a noun.
- When the second bit is 1 i.e. 0100 the word is an Adjective.
- When the third bit is 1 i.e. 0010 the word is an Adverb.
- When the fourth bit is 1 i.e. 0001 the word is a verb.
- We also have combinations like 1100 for ambiguous words that can be used both as a noun and as an Adjective.
- Another combination which we have for ambiguous words is 0110. This means that the specified word can be used both as an Adjective and as an Adverb. For pronouns we are using a separate database which contains all the possible pronouns which can be used in English Language.

B. Details of identified modules

The English sentence that is to be analyzed is given as an input by the user. The input is then sent to tokenizing function.

1) **Tokenizer**

This module generates the tokens of the given input
sentence and the delimiter that is used for tokenizing is space followed by dot(.) . It also calls the other modules when required. The tokens of the sentence are basically stored in a String array for further processing.

2) **Tagging**

The tagging module assigns tags to tokens and also search for ambiguous words and according to their type assign some special symbols to them. If we encounter words which are not present in the WordNet they are treated as unidentified. These unidentified tokens are compared with the pronoun database if these tokens are present in the

database then they are treated as pronouns. The ambiguous words are those words which act as a noun and adjective or adjective and adverb according to different context.

3) Resolving Ambiguity

The ambiguity which is identified in the tagging module is resolved using the Marathi grammar rules. These rules are:

Rule 1:

If we have a token which is assigned notation as 0110 signifies that it can be used as an adjective as well as an adverb, then such ambiguity is resolved as:

- if the next token is a noun or an adjective then the ambiguous token becomes an adjective.
- if the next token is a verb then the ambiguous token becomes an adverb.

Rule 2:

If we have a token which is assigned notation as 1100 signifies that it can be used as a noun as well as an adjective, then such ambiguity is resolved as:

- if the next token is a noun and the previous token is not a noun then the ambiguous word becomes an adjective
- otherwise it becomes an adverb.

Rule 3:

If we have a token which is assigned notation as 1100 signifies that it can be used as a noun as well as a adjective, then such ambiguity is resolved as:

- if the previous token is a noun then the ambiguous word becomes an adjective, even if the next token is a verb.
- otherwise it becomes an adverb.

4) Displaying results

This module will be displaying the final result. The tokens i.e. words in the sentences are shown with their corresponding parts of speech.

MATHEMATICAL MODEL

The fundamental equation of statistical machine translation, formulated as follows:

$$\hat{t} = \arg\max_{t} P(t|s) \qquad (1)$$

The equation expresses ˆt, the TL sentence that has the highest probability given the SL input. By means of Bayes rule this can be transformed to.

$$\hat{t} = \arg\max_{t} \frac{P(s|t) * P(t)}{P(s)} \qquad (2)$$

The denominator, P(s) is the probability of the input string which is the same for all calculations in the arg max, and can thus be disregarded. The resulting equation corresponds to the equation described at the beginning of this paragraph, where P(s|t) corresponds to the translation model and P(t) to the language model. In a way the problem has been inversed by Bayes transformed from finding the best TL string given the source, to searching for the best SL string given the target string, multiplied by the probability of this TL string

PROPOSED SYSTEM FLOW

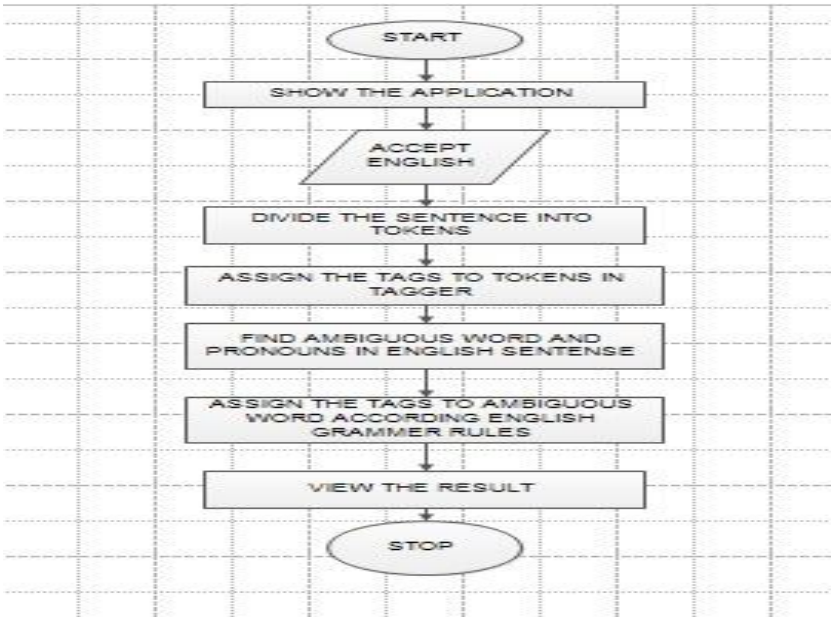
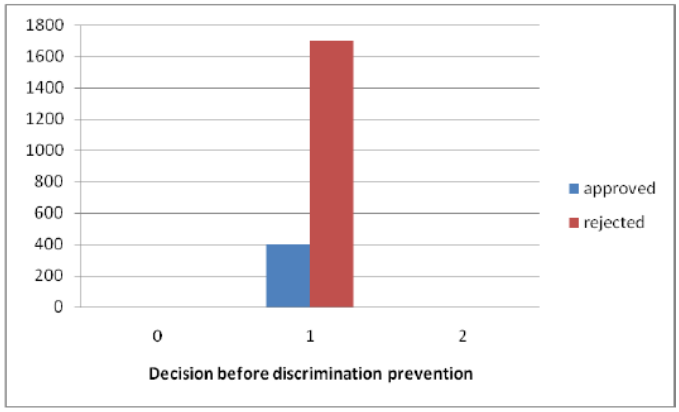


Fig 2 Flowchart of system

PERFOMANCE EVALUATION

The proposed system is effective at removing the direct and indirect discrimination from the original dataset. Anti-discrimination method was introduced in the proposed method for effective discrimination method. The existing system provides more way to the discrimination approach than the proposed system

TENTATIVE OUTPUT



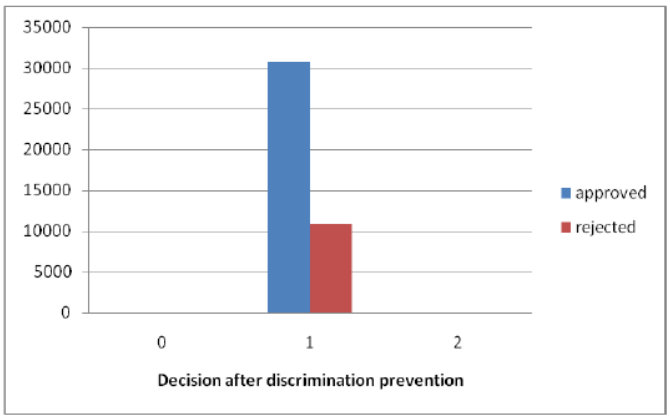


Fig 3 performance evaluation of proposed and existing System

The above figure 3 shows how the proposed and existing systems are performed. In the existing system due to the discrimination approach more number of data's are rejected in the given dataset. In proposed method Anti-discrimination approach (removal of discrimination from the original dataset by anonymzing the attributes) is introduced. Less number of dataset is rejected in the proposed method due to the effective discrimination approach.

4.CONCLUSION

In this paper we survey on Part of Speech Tagging is playing a vital role in most of the natural language processing applications. English is an ambiguous language, it is hard for tagging. The rule based POS tagger described here is resolving ambiguity and assigning the tags to the ambiguous words using English grammar rules. It provides correct tag for all the words that

are present in the WordNet. The range of words for which the POS tagger can be used, can be raised by updating the WordNet. In this paper we are going to use Natural Language Processing Approach for direct and indirect discrimination prevention. It consists of POS tagging and chunking methods. POS tagging is useful for identifying verbs, nouns, adjectives in a given line. On the basis of that we can identify the action words which may cause direct or indirect discrimination.

REFERENCES

[1] S.Hajain, J.Domingo Ferrer, and A.Martinez Balleste,"Rule protection for Indirect Discrimination Prevention in Data Mining", Proc.Eighth Int'l Conf.Modeling Decisions for Artificial Intelligence (MDAI'11).pp.211-222, and 2011.

[2] F.Kamiran,T.CaldersandM.Pecheninzkiy,"DiscriminationAware Decision Tree Learning", Proc.IEEE Int'l Conf.Data Mining (ICDM'10), pp.869-874, 2010

[3] S.Ruggieri, D.Pedreschi and F.Turini,"DCUBE: Discrimination Discovery in Databases,"Proc.ACM Int'l Conf. Management of Data (SIGMOD'10), pp, 1127-1130, 2010.

[4] D.Pedreschi, S.Ruggieri and F.Turini,"Discrimination-Aware Data Mining", Proc.14th ACM Int'l Conf.Knowledge Discovery and Data Mining (KDD'08), pp.560-568, 2008.
[5] D.Pedreschi, S.Ruggieri and F.Turini,"Integrating Induction and Deduction for Finding Evidence of Discrimination,"Proc.12th ACM Int'l Conf. Artificial Intelligence and Law (ICAIL'09), PP.157-166, 2009
[6] S.Ruggieri,D.Pedreschi and F.Turini,"Data Mining for Discrimination Discovery",ACM Trans. Knowledge Discovery from Data,vol.4,no.2,article 9,2010.P.N.Tan,M.Steinbach and V.Kumar, Introdiction to Data Mining.Addison-Wesely,2006
[7] S.Hajjain.J.Domingo-Ferrer and A.Martinez-Balleste,Discrimination prevention in Data mining for intrusion and crime detection,proc.IEEE Symp.Computational Intelligence in cyber security (CICS'11),PP,47-54,2011.
[8] Agarwal, H., Mani, "Part of Speech Tagging and Chunking with Conditional Random Fields. "In Proceedings of NLPAI Machine Learning Workshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad,India (2006).
[9] Pattabhi, R.K.R., SundarRam, R.V., Krishna, R.V., Sobha, L.,"A Text Chunker and Hybrid POS Tagger for Indian Languages"In Proceedings of International Joint Conference on Artificial Intelligence Workshop on Shallow Parsing for South Asian Languages, IIIT Hyderabad, Hyderabad, India (2007).
[10] Fahim Muhammad Hasan," Comparison Of Different Pos Tagging Techniques", Brac University, Dhaka, Bangladesh, , pages: 13,2006.
[11] Ekbal, A., Mandal, S.: POS Tagging using HMM and Rule based Chunking. In: Proceedings of International Joint Conference on Artificial Intelligence Workshop on Shallow Parsing for South Asian Languages, IIIT Hyderabad, Hyderabad, India (2007).
[12] Awasthi, P., DelipRao, Ravindran, B.: Part of Speech Tagging and Chunking with HMM and CRF. In: Proceedings of NLPAI Machine LearningWorkshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad, India (2006).
[13] Baskaran, S.: Hindi Part of Speech Tagging and Chunking. In: Proceedings of NLPAI Machine Learning Workshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad, India (2006).
[14] Karthik Kumar G, Sudheer K, Avinesh Pvs, "Comparative Study of Various Machine Learning Methods For Telugu Part of Speech Tagging", In Proceedings of the NLPAI Machine Learning 2006 Competition. Pallavi Bagul et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1322-1326