Identifying the presence of sexism in workplace comments

Final Presentation

Ishika Joshi
Keshav Gambhir
Sachleen Kaur
Shrivatsa Mishra

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Motivation

- Sexism at work is **discrimination** based on a **person's sex** that occurs in a place of employment. Here, we are focusing on sexist comments targeted at **women employees**.
- Sexism is prevalent in the form of **formal/informal remarks** and communication, **lack of professional treatment** (eg. promotions), and many such areas with varying trivialities which makes the work environment really **unhealthy** and **demotivating** for the **women workforce**.
- This is confirmed by the official statistics that show even in developed countries like the USA, **42% of the women** face sexism at workplace versus **22% for men**. This is why we decided to narrow down our interests to sexism faced by women.

Motivation

- It is reported that **77% women** have at least once faced sexism in forms of speech. This is all the more prevalent in workplaces. They are often looked down upon and demeaned due to their gender.
- The frequency of this occurrence made us take up sexist comments at workplace. We believe this domain caters to our interest in taking up and serving grounded problems that can make a real impact. Hence, we are highly motivated to move forward with this theme.

Literature Review

We looked at 3 different papers, each looking at the same topic using different methods. From these we hoped to identify which method would give us the best accuracy along with an expected accuracy. These also included some complex NLP models which were not in our course so we chose to ignore them.

- **Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data**: This paper has a data-set of Spanish sexist speech on Twitter and used that to compare traditional methods and neural-network-based algorithms.

  They compared methods based on **Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)** to Bidirectional Long Short-Term Memory(BiLSTM) and Bidirectional Encoder Representations from Transformers(BERT). The results showed that BERT outperforms the rest by a tiny margin. Both SVM's and BERT performed similarly with marginal differences having precisions of 0.61 and 0.62 respectively. **The best model that we should use according to this is a SVM.**

Literature Review

- **Automated Hate Speech Detection and the Problem of Offensive Language**: This paper attempted to identify hate speech online through the use of **tf-idf SVM**.

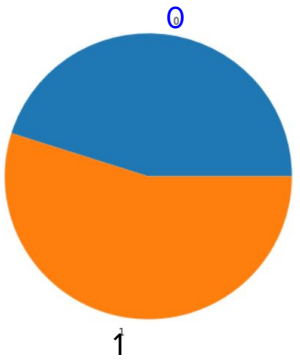
  They tried multiple different models(**logistic regression, naive Bayes, decision trees, random forests, and linear SVMs.**) however this gave the best results. Their model achieved a precision of 0.91 however seemed to misidentify many, 31%, hate comments as simply offensive.

- **When does a Compliment become Sexist? Analysis and Classification of Ambivalent Sexism using Twitter Data** :This paper identifies the border between compliments and sexist compliments through the use of different models.

  They tried **tf-idf SVM, Sequence to Sequence,** and **FastText** models. Their **SVM** model had accuracy of 0.97, 0.87, 0.80 for benevolent, hostile, and others after 10 fold validation. SVM also seems to work better on benevolent comments as compared to hostile comments.

Understanding the Dataset

- The dataset comprises of over 1,100 examples of statements or comments passed in workspaces.
- These statements are labelled as 1 or 0. The label ‘1’ denotes a sexist statement and 0 denotes ambiguous or neutral cases.
- 55% of the dataset includes statements of ”benevolent” sexism procured from twitter, other sources of workplace-related sexist speech are included to keep the source contexts of the workplace statements diversified in order to reduce overfitting on confounding keywords and phrase constructions.

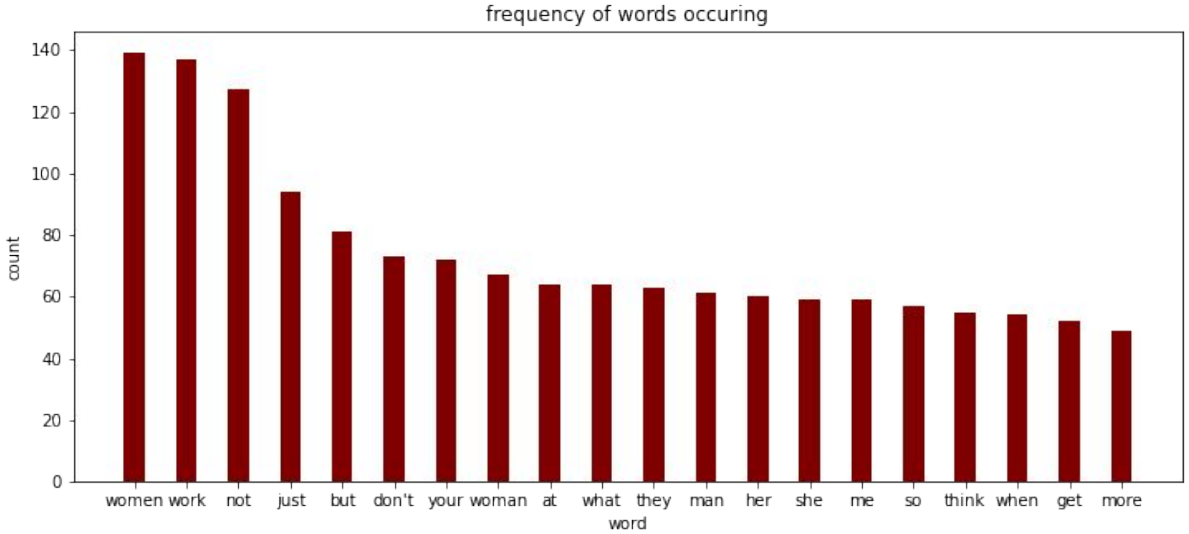


0 : Ambiguous/Neutral (513)
1 : Sexist Statements (624)

Hence, the dataset is not biased.

- For the dataset we have, we will have to pre process it to perform cleaning, get rid of unnecessary noise, lemmatize the data, etc. Hence, some amount of data preprocessing has been carried out.

Understanding the Dataset




```
women :  139
work :  137
not :  127
just :  94
but :  81
don't :  73
your :  72
woman :  67
```

To ensure the relevancy of the dataset to our objective, we tried to plot the occurrences of the most occurred words in the entire dataset. We performed some text cleaning and removed the stop-words from the dataset. As we can see, the results we obtained seem to relevant to sexism against women at workplaces from the looks of it.

Methodology: Data Preprocessing

To preprocess the data, we performed some text cleaning. This was necessary to make sure all the textual inputs are uniform and no difference in output could be caused due to textual trivialities.

The following are the operations we performed to text:

- **Contractions:** Contractions have been replaced by their full forms. This is done to obtain text standardization.
- **Remove special characters:** All characters like comma, semicolon, etc have been removed as we do not particularly need them in our context.
- **Lemmatization:** We have grouped together words in inflicted forms. To do so, we have used Lemmatizer from WordNet.

Methodology: Data Preprocessing

- **Lower case:** We lowered case all the words to make sure capital initials won't make a difference in the way the data is perceived. This is because we do not need to treat 'women' and 'Women' differently in our context.
- **Tokenization:** We have split the text statements into smaller groups or chunks. This will aid in processing it. To split it, we have used WordPunctTokenizer from the nltk library.
- **Stopwords:** Some words have been treated as stopwords. These are conjunctions or articles which are not contributing any information to the model and will only create noise. To do so, the ntlk package is used.

Methodology: Feature Extraction

- ➔ Machine Learning models don't work with text as the input data
- ➔ Two methods of converting statements into numerical vectors
  - ◆ **BoW (Bag of words):** Based on the word count. The order of the words does not matter.

    Bow (w, d) = Number of times word w appears in document d

  - ◆ **Tf-Idf(term frequency-inverse document frequency):** Not just based on the word count, but more on the normalised count where each word is categorised by the number of documents it occurs in.

    Tf-idf(w, d)= Bow(w, d) * log(Total Number of Documents /(Number of documents in which word w appears))

- ➔ For this project we have focused more on the Tf-Idf extraction technique

Methodology: ML Models

After feature engineering, we used various machine learning models to train the dataset. We used a train-test split of 7:3 for all cases.

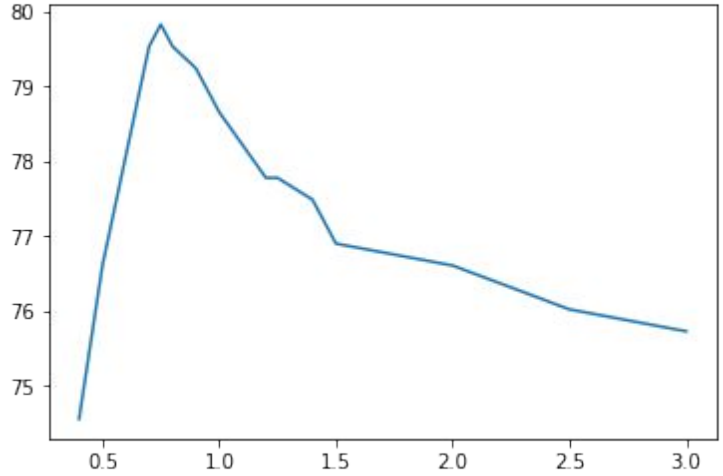
1. **Logistic Regression:** Used for binary classification. Here the input features form a weighted combination, the sigmoid value of which determines if the output is 0 or 1. This model was trained on both BoW and Tf-Idf extraction techniques.
2. **Naive Bayes:** This is commonly used on for text classification as we obtain the probability of each word through BoW and tfidf allowing us to use Naive easily. Here we have used tf-idf only as it works better than bow in most cases.
3. **Random Forest:** Random forest is an ensemble method that uses decision trees to give out predictions. It constructs various decision trees to give out results for the tasks of classification and regression. In the project we have used Tf-Idf technique for feature extraction and the output is served into a random forest classifier to predict the output.

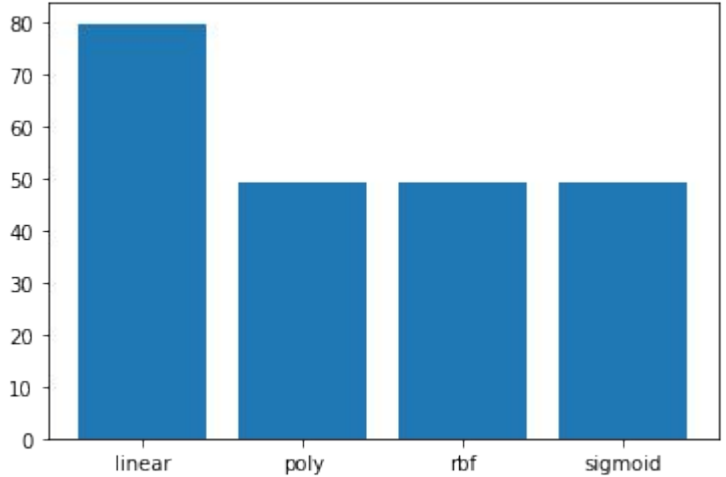
Methodology: Support Vector Machine

- Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression tasks.
- SVM uses the idea of plotting the points in an n-dimensional plane and then separating them with a hyperplane. The hyperplane is used to differentiate between the various points in n-dimensional space. There are mainly two types of hyperplanes: Linear and non-linear hyperplanes. In linear hyperplanes, points are linearly separable; they can be classified using a single straight line.
- The classifier used to separate this type of data is called linear SVM. Non-Linear SVMs are used to separate non-linear data, that is, the data that cannot be separated using a straight line.
- SVM algorithm helps to find the best line or decision boundary. Then it finds the closest points of the lines from both the classes. These points are called support vectors, and the distance between the vector and the hyper plane is called the margin. The goal of the SVM is to maximize the margin found by the SVM. The hyperplane with the maximum margin is called an optimal hyperplane. SVMs are highly effective in multidimensional spaces and when the number of dimensions is more significant than the number of samples.

Methodology: Support Vector Machine

For the SVM we attempted to find the optimal values for C, the regularization parameter as well as find the type of kernel we should use. For this we ran the dataset through the SVM with different hyperparameters and stored their accuracies. From this we learned that it is much better to change the value of C from the default 1 to a value of 0.75 as it offered the best accuracy. We also found that a linear kernel worked best for us as it gave a much better accuracy as compared to the others.





Methodology: Introducing ambiguity label

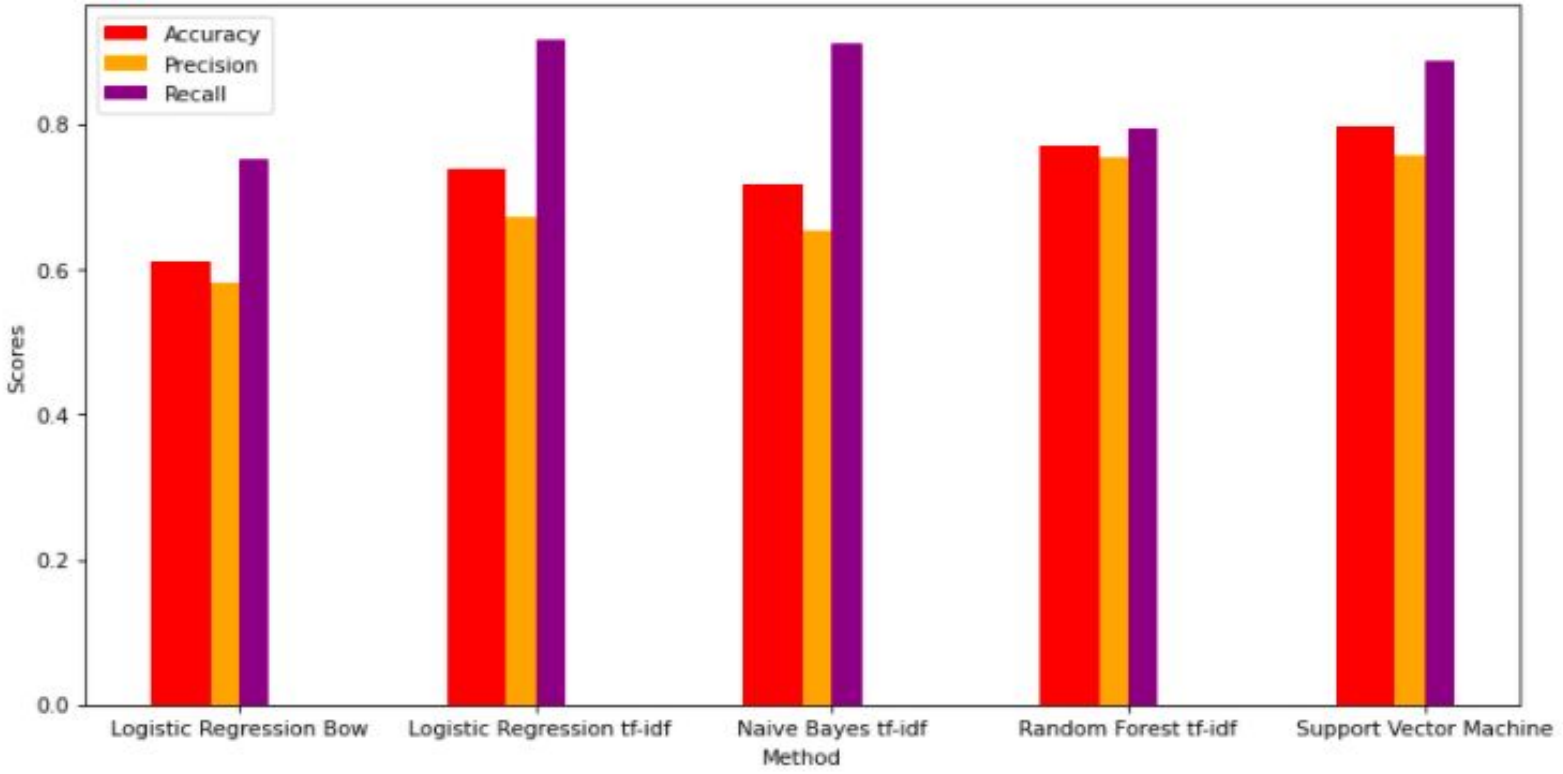
Since the dataset had the final output variable as binary classification, i.e. classifying each statement as sexist or not sexist, we realised how the model might not always give us accurate results.

In a real life scenario, sexism is a rather sensitive topic and labelling a statement sexist or not sexist without taking the complete context can be quite extreme. So, we decided to introduce a new output label in our model. This can be labelled as an ambiguous' classification of a statement. The model will present such a classification when the statement does not exactly classify as a sexist or a non sexist statement. We take the probabilities using predict_probability and classify them as 1 or 0 only if the difference b/w their probabilities is large enough, if not then we classify it as 0.5 or ambiguous.

Results

Logistic Regression Bow	Logistic Regression tf-idf	Naive Bayes tf-idf	Random Forest tf-idf	Support Vector Machine tf-idf
Accuracy: 0.6111111111111111 **Precision:** 0.5825688073394495 **Recall:** 0.7514792899408284	**Accuracy:** 0.736842105263158 **Precision:** 0.670995670995671 **Recall:** 0.917159763313609	**Accuracy:** 0.716374269005 **Precision:** 0.652542372881356 **Recall:** 0.9112426035502958	**Accuracy:** 0.7690058479532164 **Precision:** 0.7528089887640449 **Recall:** 0.7928994082840237	**Accuracy:** 0.79789799 **Precision:** 0.7573689 **Recall:** 0.887573

- Through the above tables we can see that Random Forest allows us the best result giving an accuracy of around 76%. Logistic Regression with BoW on the other hand is the worst, with an accuracy of 61%. Logistic Regression and Naive Bayes performs relatively well on the data however is still prone to error as compared to the random forest. However SVMs offer the best accuracy of the bunch achieving almost 80% accuracy.
- Between BoW and tf-idf feature extraction using Logistic Regression, Bow appears to be performing much worse, giving us an accuracy of 61\% as compared to the 73\% of tf-idf. Thus we can assume that tf-idf is the superior method of feature extraction and we have use that in the following algorithms.



Visualisation of the results obtained from various ML models

Analysis

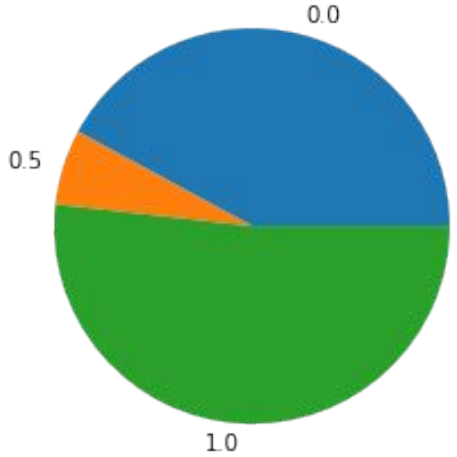
- We have used the SKLearn model for logistic regression which allows us L1 and L2 regularisation along with the normal regression. This allows it to work well on a dataset with a large number of features, many of which might not be relevant. This is because tf-idf causes a lot of noise in the data due to giving importance to each word. Still the large number of variable as compared to the data causes logistic regression to not work as well.

- Naive Bayes also works well but not as well as random forest. This is because it assumes all the variables are independent of each other, which is not the case for text data, two words on their own may mean something totally different as compared to when they are used together. Hence Naive Bayes does not work as well on text type data due to this.

- Random Forest works better than the previous two, with Logistic Regression coming close thanks to it taking different random features for each decision tree, it is able to work on eliminating redundant features quite well.

Analysis

- The best accuracy is observed when SVM is used with linear kernel. The accuracy score is approximately 0.797 with precision and recall equal to 0.7528 and 0.8828 respectively. As the literature review also signifies the SVM would give the best accuracy. This hypothesis has been verified by our model as well. The accuracy observed was better than any other model that is logistic regression, naive bayes and Random forest.



Pie chart to visualise the number of number of sexist, non sexist and ambiguous statements as given in the test set

Conclusion

- We have attempted various Machine Learning algorithms in order to obtain the best model. We obtained the best results from Support Vector Machine according to the literature we have read.
- We also added the ability to detect ambiguity, this is done through comparing the probabilities of being sexist or not, if the difference is not significant enough it classifies the statement as ambiguous, this is done as some statements are not clear cut. This feature allows us to not make error is such crucial cases. This is extremely important as cases of sexism are a serious matter and so we must be sure of the result and if uncertain, say so.
- Through this project we have used a dataset of labeled statements and multiple different machine learning models to classify sexist statements. Topics such as these are extremely important to not only identify but call out sexism at workplace. We have hoped to do it justice through this exploration on the different ways to identify them. In the future this project has the capability to be deployed to help combat sexism in the workplace.

Individual Contributions

- **Literature Review** - Shrivatsa Mishra & Keshav Gambhir
- **Data Visualization** - Ishika Joshi & Sachleen Kaur
- **Preprocessing** - Ishika Joshi & Keshav Gambhir & Sachleen Kaur & Shrivatsa Mishra
- **Feature Extraction** - Sachleen Kaur & Shrivatsa Mishra
- **Logistic Regression(bow)** - Ishika Joshi
- **Logistic Regression(tf-idf)** - Sachleen Kaur
- **Naive Bayes Model** - Shrivatsa Mishra
- **Random Forest** - Keshav Gambhir
- **Support Vector Machine Mode**l - Shrivatsa Mishra & Sachleen Kaur & Ishika Joshi & Keshav Gambhir
- **Analysis Visualisation** - Ishika Joshi & Sachleen Kaur
- **Hyperparameter Analysis** - Shrivatsa Mishra & Keshav Gambhir
- **Ambiguous Model Labeling** - Shrivatsa Mishra & Sachleen Kaur & Ishika Joshi & Keshav Gambhir
- **Results and Analysis** - Shrivatsa Mishra & Ishika Joshi

References

[Chatterjee, 2018] Chatterjee, R. (2018). A New SurveyFinds 81 Percent Of Women Have Experienced SexualHarassment.NPR.

[Davidson et al., 2017] Davidson, T., Warmsley, D., Macy,M., and Weber, I. (2017).Automated Hate SpeechDetection and the Problem of Offensive Language.arXiv:1703.04009 [cs]. arXiv: 1703.04009.

[Ismiguzel, 2021] Ismiguzel, I. (2021).Applying TextClassification using Logistic Regression: A comparisonbetween BoW and Tf-Idf.

[Javapoint,] Javapoint. Support Vector Machine (SVM)Algorithm - Javatpoint.

[Jha and Mamidi, 2017] Jha, A. and Mamidi, R. (2017).When does a compliment become sexist? Analysis andclassification of ambivalent sexism using twitter data.InProceedings of the Second Workshop on NLP andComputational Social Science, pages 7–16, Vancouver,Canada. Association for Computational Linguistics.

[KOWALCZYK, 2017] KOWALCZYK, A. (2017). SVMs- An overview of Support Vector Machines.

[Parker,] Parker, K. 42% of US working women havefaced gender discrimination on the job.

[Rodr´ıguez-S´anchez et al., 2020] Rodr´ıguez-S´anchez, F.,Carrillo-de Albornoz, J., and Plaza, L. (2020). Auto-matic Classification of Sexism in Social Networks: AnEmpirical Study on Twitter Data.IEEE Access

Thank You!!