AMERICAN UNIVERSITY OF BEIRUT

BENEVOLENT SEXISM DETECTION IN TEXT: A DATA-CENTRIC APPROACH

by
ZAHRAA JIHAD BERJAWI

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science in Business Analytics
to the Suliman S. Olayan School of Business
at the American University of Beirut

Beirut, Lebanon
September 2022

AMERICAN UNIVERSITY OF BEIRUT

BENEVOLENT SEXISM DETECTION IN TEXT: A DATA-CENTRIC APPROACH

by

ZAHRAA JIHAD BERJAWI

Approved by:

Dr. Wael Khreich, Assistant Professor — Advisor
Suliman S. Olayan School of Business

Dr. Wissam Sammouri, Assistant Professor of Practice — Member of Committee
Suliman S. Olayan School of Business

Dr. Sirine Taleb, Lecturer — Member of Committee
Suliman S. Olayan School of Business

Date of thesis defense: September 13, 2022

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: Berjawi Zahraa Jihad

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

☒ As of the date of submission

□ One year from the date of submission of my thesis.

□ Two years from the date of submission of my thesis.

□ Three years from the date of submission of my thesis.

Zahraa September 15, 2022

Signature Date

ACKNOWLEDGMENTS

This thesis would not have been possible without the support of many people. First, many thanks to my thesis advisor Dr. Wael Khreich, who introduced all the concepts of machine learning to me, guided, supported, and helped me in making a lot of sense out of the confusion.

Also, special thanks to Dr. Wissam Sammouri, the director of the MSBA program at AUB, for his invaluable inspiration, sincerity, and encouragement.

I am grateful for my parents and my brothers, whose constant support kept me going. I owe my deepest gratitude to my fiancé for his unconditional love and motivation throughout the thesis process and every day.

Finally, I wouldn’t have been here today without the Middle East Partnership Initiative (MEPI) scholarship, which funded my master’s degree from the beginning to the very end.

ABSTRACT
OF THE THESIS OF

Zahraa Jihad Berjawi for Master of Science in Business Analytics

Title: Benevolent Sexism Detection in Text: A Data-Centric Approach

The Ambivalent Sexism theory divides sexism into two-dimensional ideologies: benevolent sexism and hostile sexism. Hostile sexism has been associated with short-term harmful impacts, and benevolent sexism has been proven to have more severe long-term effects on women's well-being, their representation, and gender equality in societies. Recently, research has been directed toward the detection and mitigation of hostile sexism, and minimal efforts have been done with the aim of detecting and mitigating benevolent sexism. Adversely, since benevolent sexism is associated with a seemingly positive expression, detecting and mitigating its online spread is a challenge that needs the attention of social scientists, gender scholars, and data scientists. In this paper, we aim toward creating a benevolent sexism detection system. To the best of our knowledge, the research area lacks a representative benevolent sexism dataset. Thus, to be able to train supervised machine learning models, we collected and labeled a dataset of benevolent, hostile, and non-sexist statements collected from quotes' websites, online articles, and the Google Advanced Search tool. Further, we trained several machine learning models and incrementally tuned and optimized the best classifier for the detection of benevolent sexism. Then, we validated our model's performance on similar and broader context datasets and detailed its strengths, weaknesses, and areas of improvement. Our final results confirm our model's ability to detect benevolent sexism in a generalized context. To emphasize, the dataset collected was proven to perform well in the representation of the benevolent sexism expression. In conclusion, this research is a steppingstone to creating a self-learning, data-centric benevolent sexism detection system.

TABLE OF CONTENTS

ILLUSTRATIONS

Figure

TABLES

Table

ABBREVIATIONS

TF-IDF	Term Frequency- Inverse Document Frequency
LR	Logistic Regression
SVM	Support Vector Machines
SVC	C-Support Vector Classification
DT	Decision Tree
RF	Random Forest
GNB	Gaussian Naïve Bayes
CNB	Complement Naïve Bayes
MNB	Multinomial Naïve Bayes
TP	True Positives
FP	False Positives
TN	True Negatives
FN	False Negatives
TPR	True Positive Rate
FPR	False Positive Rate
ROC	Receiver Operating Characteristic Curve
AUC	Area Under the ROC Curve
Cross-Val	Cross Validations
G-Mean	Geometric Mean

CHAPTER 1

INTRODUCTION

Sexism is defined as the discrimination that is based on gender or sex, and its concept has been first linked to raising awareness of the oppression of women and girls around the 1960s (Masequesmay, 2021; Meriam-Webster, n.d.). More specifically, Glick and Fiske (1996) argue that sexism is not always characterized by prejudice against women. Instead, they present the Ambivalent Sexism Theory in which they view sexism as a multidimensional concept that encompasses two types: Benevolent Sexism and Hostile Sexism (Glick & Fiske, 1996). Further, they suggest that these two components have three subtypes: Paternalism, Gender Differentiation, and Heterosexuality (Glick & Fiske, 1996). Further, research on the impact of sexism against women argues that benevolent sexism has more long-term effect on societies, as it restricts women to specific roles and undermines their value through a positive tone of expression (Barreto & Ellemers, 2005; Glick & Fiske, 1996).

Recently, research has been mainly focused on the detection and identification of the hostile form of sexism. Their research approaches included building a corpus of hostile sexist statements and training different machine learning and/or deep learning models for the classification of these statements (Waseem & Hovy, 2016: Waseem, 2016). Similarly, there has been wide research focus on the extreme form of hostile sexism, which includes misogyny and sexual harassment (Anzovino, Fersini, & Rosso, 2018: Fersini, Rosso, & Anzovino, 2018: Fersini, Nozza, & Rosso, 2018).

However, little research has been focused on the benevolent aspect of sexism. For instance, Jha and Mamidi (2017) constructed a dataset of benevolent sexist tweets and used pre-existing hostile sexist tweets (Waseem & Hovy, 2016) to build a classifier that helps in detecting sexism in its two forms. However, since the pre-existing benevolent sexism data (Jha & Mamdi, 2017) depends on the availability of tweets and retweets, more than 92% of the tweets were lost after extraction and cleaning.

To the best of our knowledge, the research community still lacks a reliable and representative dataset to be used in the automatic detection of benevolent sexism. This deficiency can be due to several reasons. One of them would be the seemingly benign form of benevolent sexism. This makes detecting such statements challenging, as the exposed person might not be able to distinguish benevolent sexism from an appealing complimentary conversation. In addition, the ability to detect gender discrepancies requires knowledge of gender issues. This task can also be time-consuming since it needs verification from different ends to minimize the factor of subjectivity in the data collection process.

Accordingly, this research aims at contributing to this field through a data-centric approach to benevolent sexism detection. In the aim of constructing this system and training supervised machine learning models, we first collected a representative sexism dataset with statements labeled as Benevolent Sexism, Hostile Sexism, or Non-Sexism. Then, using the collected data and techniques of NLP and machine learning, we trained and incrementally optimized machine learning models that will allow the detection of potential benevolent sexist statements out of large amounts of data.

Following our modeling experimentation, the Support Vector Classification (SVC) model yielded the best results and was used for the detection of benevolent sexist statements. The model was tested in a broader context, and the results showed a high generalizability performance. This indicates that our dataset is representative of the embedded benevolent sexism expression and that our model can be executed into a benevolent sexism detection system for broader use.

CHAPTER 2

LITERATURE REVIEW

2.1. Benevolent Sexism Against Women

In this section, we explain the manifestation of gender stereotypes and their expression in language. Further, we detail the multidimensionality of sexism (Hostile and Benevolent), elaborate on each type's subcategories (Paternalism, Gender Differentiation, and Heterosexuality), and state their direct and indirect impacts on different aspects.

2.1.1. Gender Stereotypes in Language

Gendered grammatical conventions utilized in language may be highly associated with the manifestation of gender norms and the formation of gendered stereotypes (Prewitt-Freilino, Caswell, & Laakso, 2012). In other words, these stereotypes construct the societal belief system on what is expected out of each gender, and they are manifested in a biased language that is used to maintain the notion of "opposite" sexes with distinct responsibilities (Cameron, 2003; Stahlberg et al., 2007). Among the beliefs that constitute these societal biases are women being kind, emotional, and more inclined towards participating in the domestic sphere (UNODC, 2018). On the other hand, men are depicted as more fit to engage in public life for their courage, independence, and leadership traits (UNODC, 2018). As a result, these stereotypical beliefs would be reflected in the everyday lexical choices that refer to men or women, including prejudice or stereotypes that are based on gender or, in other words, sexism (Menegatti & Rubini, 2017).

2.1.2. Sexism and Ambivalence

Sexism is defined as discrimination that is based on gender or sex (Merriam-Webster, n.d.). Women's statuses have dramatically changed since the introduction of this concept in the 1960s, and sexism has transformed into different forms that might not always reflect prejudice and hatred towards women (Swim & Hyers, 2021; Mills, 2008).

More specifically, Glick and Fiske (1996) argue that sexism is not always characterized by prejudice against women. Instead, they present the Ambivalent Sexism Theory in which they view sexism as a multidimensional concept that encompasses two types: benevolent sexism and hostile sexism (Glick & Fiske, 1996). The hostile sexism form of sexism is expressed in a negative, blatant, and aggressive manner, and it reflects men's hatred toward women (Glick & Fiske, 1996). On the other hand, the seemingly benign form of sexism is expressed in a chivalrous tone, as it explains men's dominance through their affection and love for women (Glick & Fiske, 1996). At the core of the Ambivalent Sexism Theory, the two types of sexism are composed of three shared sub-categories that are paternalism, gender differentiation, and heterosexuality.

Table 2. 1

Hostile Sexism Sub-Categories

Subtype	Dominative Paternalism	Competitive Gender Differentiation	Heterosexual Hostility

Definition	It explains patriarchy as the need for a superior male figure above women since they are viewed as not being fully competent.	It justifies men's dominance by perceiving them as having the important traits and power to head social institutions.	It is accompanied by hostility toward women under the belief that women use their sexual characteristics to deceive and dominate men.
Example	'Women should stay at home and men should do the work'	'The people at work are childish. It's run by women and when women don't agree to something, oh man'	'Hate these blonde bitches already'

Table 2. 2

Benevolent Sexism Sub-Categories

Subtype	Protective Paternalism	Complementary Gender Differentiation	Heterosexual Intimacy
Definition	It is the belief that women are the weaker sex, and that men should provide them with protection, love, and cherish.	It views women as having soft and positive traits that complement those of men. It also uses the physical differences between the sexes to justify the "dyadic dependency of men and women".	It explains that men's dependence on women may be related to their sexual motivation and a genuine psychological need for their closeness.
Example	'A real woman can do it all by herself but a real man won't let her.'	'It's so good that I thought your brother wrote it!'	'What is man without the love of a woman!'

2.1.3. Impact of Benevolent Sexism

Despite the pleasant feelings indicated by the perceiver in most cases, benevolent sexism restricts women to specific roles and undermines their value through a positive tone of expression (Glick & Fiske, 1996). More specifically, benevolent sexism has various detrimental effects on women, relationships, and society. This is because the benevolent sexism subtypes suggested by Glick and Fiske (1996), which are protective paternalism, complementary gender differentiation, and heterosexual intimacy can be interpreted and manifested in diverse ways.

2.1.3.1. Protective Paternalism

- Women's Well-Being

Four experiments were conducted by Dardenne, Dumont, and Bollier (2007) where female participants were tested for an imaginary job position that requires feminine characteristics. In their first experiment, the managers explained the tasks to women in one of three different manners, nonsexist, hostile sexist, and benevolently sexist (specifically protective paternalism). Precisely, the construction of these distinct sexist forms was driven by the Ambivalent Sexism Theory suggested by Glick and Fiske (1996). In the hostile sexist explanation, women were referred to as the weaker sex and men as the stronger while in the benevolent sexist explanation, women were referred to as the nice sex who needs help from men in their tasks. The impact of these different instances on women was tested through their performances in the assigned tasks. Further, the second experiment was similar to the first; however, the impact of the three different instances (non-sexist, hostile sexist, and benevolently sexist) was tested by asking women about their motivation toward the completion of the tasks (Dardenne, Dumont & Bollier, 2007). In the third experiment, participants were given written instructions that included either non-sexist, protective paternalism, or complementary gender differentiation. In detail, protective paternalism in the instructions stated that women will be working with men only, and this should be a good distribution since men will help them in their work. Additionally, complementary gender differentiation was stated in the manner that women will be working with men only, and this is a good distribution since the firm needs women's refined characteristics that are lacking in men. Moreover, the fourth experiment included participants' exposure to expressions of one of the two hostile or benevolent sexist written instructions (Dardenne, Dumont & Bollier, 2007). The instructions and procedures were similar to that in the first and the second experiments in addition to measuring the participant's confidence, self-doubt, and preoccupation while solving the tasks. As a result, after measuring the impact of these instances on the working memory of women, it was concluded that women who are subject to notions of protective paternalism experience mental intrusions that affect their concentration on their tasks, thus, slowing down their performances and increasing their feelings of incompetency (Dardenne, Dumont & Bollier, 2007).

Further, consistent with the previous study, Oswald, Baalbaki, and Kirkman (2018) conducted two survey studies, each on a varied set of women participants. The surveys include sexist incidents of the three benevolent sexism sub-categories in addition to hostile sexism that might have occurred with the participants. Afterward, other surveys were completed by participants to measure the impact of these incidents on women's self-doubt (Oleson et al. 2000; Mirels et al. 2002), self-esteem (Rosenberg, 1965), perceived life satisfaction (Cummins et al. 2003), psychological flourishing (Diener et al. 2010), behavior with authority figures (Rigby 1987), and their endorsement of sexist attitudes (Glick & Fiske, 1996, 1997). In this examination of females' experiences of benevolent sexist behaviors of distinct categories, protective paternalism was rated as the most stressing subcategory of benevolent sexism, as it mostly impacts women’s well-being and self-concept (Oswald, Baalbaki, & Kirkman, 2018).

- Gender Equality

Moya et al. (2007) conducted three studies to assess women’s endorsement and reactions to protective paternalism. Women and men were presented with both hypothetical and realistic scenarios that offer women a possibly dangerous activity, and their responses to hostile and protective benevolent rejection from their co-workers and partners were analyzed (Moya et al., 2007). For instance, in one of their studies, a community sample of women was exposed to a hypothetical scenario in which a man decides to take the responsibility of driving them on a long trip. The justification of the male actors’ actions varied between hostility and protection, and an actor was either a coworker or a husband. This study concludes that when notions of protective paternalism are expressed with affection, women face difficulty in differentiating men’s chivalry from their attempts to manipulate and control their behavior. Thus, women are often willing to accept men’s restrictions and dominance by justifying them as acts of love and care (Moya et al., 2007). More specifically, Shnabel et al. (2016) focused on assessing females’ dependence on males in doing difficult tasks and males’ willingness to provide “dependency-oriented help”. The results were consistent with previous studies, indicating that protective paternalism poses danger to gender equality by discouraging women from seeking their independent success and relying on their male partners to assist in their achievements (Shnabel et al., 2016: Viki, Abrams & Hutchison, 2003).

- Women’s Representation

In addition, King et al. (2012) conducted five survey studies to examine the potential gender discrimination in the quality and quantity of advancing work experiences. The first two studies were conducted on female and male managers and healthcare employees. Participants were asked to choose among a set of developmental experiences that they have done with diverse levels of difficulty. Then, the remaining three experiments focused on asking participants about the preferred challenge level at their work and whether they would assign complex tasks for females or males in an imaginary management position (King et al., 2012). Further, they used the Ambivalent Sexism Inventory to measure the participants' endorsement of benevolent sexism (Glick & Fiske, 1996, 1997). In conclusion, King et al. (2012) assert that the manifestation of the protective paternalism component of benevolent sexism in the workplace has a tremendous impact on the underrepresentation of women in the workplace. These findings were due to the fact the managers most likely offer challenging tasks to men, believing that they should "protect" women from these inconvenient situations (King et al., 2012). This, in return, made women stay in their positions for a longer time while men are getting promoted (King et al., 2012).

2.1.3.2. Complementary Gender differentiation

- Women's Well-Being

Dardenne, Dumont, and Bollier (2007) emphasize that benevolent sexism expressed through gender differentiation alone had a significant impact on women's cognitive performance. Their study was done on female undergraduates and tested the participants for a job application where "feminine" characteristics are required. Comments on the job requirements were presented in one of three forms, hostile sexism, benevolent sexism (complementary gender differentiation), and non-sexism (Dumont, Sarlet & Dardenne, 2008). The autobiographical memories of the participants were analyzed after the exposure to benevolent and hostile sexism by testing their ability to memorize specific sentences and directly filling out surveys about their reaction to the job requirements. Accordingly, Dumont, Sarlet, and Dardenne (2008) show that women's exposure to benevolent compliments on their gender group often affects their self-construal and leads to cognitive incompetence.

- Gender Stereotypes

Fields, Swan, and Kloos (2010) study women's responses to their opinions on being "a woman" by assessing the impact of women's exposure to benevolent sexism from their community. In their study, young adult women were asked to draft an essay on how their thoughts about being a woman were influenced by those of their grandmothers and mothers. They conclude that women are most likely to adopt benevolent sexist attitudes and beliefs that hold rewards for them due to their gender differentiation from men (Fields, Swan & Kloos, 2010). Thus, the impact of gender differentiation on women might not always be a result of an external influence, but it might stem from their adopting sexist beliefs through various stages in life. As a result, the adoption of these beliefs in both men and women may lead to rape myth acceptance as a justification for women's violation of these gender stereotypes and exposing themselves to sexual attack (Chapleau, Oswald & Russell, 2007). This finding was based on a survey study that measured the correlation between rape myth acceptance and the endorsement of benevolent sexist thoughts in female college students (Chapleau, Oswald & Russell, 2007).

Additionally, Jost and Kay (2005) conducted several experimental studies on men and women to measure the impact of complementary gender stereotypes on maintaining the current gender system. In their first study, participants were asked to indicate what is suited more for men or women between several communal or agentic traits using a 1-10 scale. The next study included splitting the participants into two groups, where the first was asked to indicate the level of agreement with a stereotypical statement on a 0-5 scale, and the other group was assigned to state the degree of ambiguity of the statements' wording on a 0-5 scale. In the third study, participants were exposed to two statements of research and then asked whether men or women would make better managers. The first context indicated that communal traits are suggested by research as important managerial skills, and the second context indicated that agentic traits are more important in managers (Jost & Kay, 2005). After completing a questionnaire that measures their degree of acceptance of the current gender-specific system, the analysis concluded that benevolent gender differentiation is more likely to be embraced by women who see an advantage in maintaining the current system with its gender inequalities (Jost & Kay, 2005). Similarly, Barreto and Ellemers (2005) conducted a survey study to measure the degree of similarity between men and women in their perception of benevolent sexism as prejudicial. They argue that hostile sexism is more rejected than benevolent sexism by both men and women (Barreto and Ellemers, 2005).

- Gender Equality

The endorsement of benevolent gender stereotypes entails a greater impact on gender inequality and the justification of the gender-specific system (Becker & Wright, 2011; Barreto & Ellemers, 2005). To illustrate, Becker and Wright (2011) conducted four web-based experiments on prospective teachers, psychology, and other students. They measured women's collective action intentions and the impact of women's justification of the gender-specific system. For instance, in one of their experiments, female psychology students were exposed to benevolent sexism, hostile sexism, gender-neutral, and gender-unrelated instances. Also, the female students' intentions for participating in collective actions were measured by offering women the opportunities to participate in activities that support women's rights. Their study asserts that exposure to gender differentiation decreases women's participation in collective action against gender inequality since they perceive the gendered system as advantageous (Becker & Wright, 2011).

2.1.3.3. Heterosexual Intimacy

- Relationship Expectations

Benevolent sexism in heterosexual relationships might appear as early as the stages of partner selection (Chen, Fiske & Lee, 2009). In other words, men who endorse benevolent sexism choose women who have submissive traits, such as being home oriented. Also, benevolent sexist women prefer a man with more dominant characteristics (Chen, Fiske & Lee, 2009). To illustrate, Chen, Fiske, and Lee (2009) conducted a survey study on undergraduate and graduate students who had been involved in committed relationships. Their survey included testing the degree of endorsement of benevolent sexism among participants in addition to the participants' opinions on power-related gender-role ideologies. Similarly, Lee et al. (2010) conducted a two-part survey study on university students where one is intended to collect participant's beliefs regarding relationships ideals with their opposite gender, and the second measured the participants' endorsement of benevolent sexism (Chen, Fiske & Lee, 2009). In conclusion, both studies indicate that benevolent sexism is embedded in the context of heterosexual relationships in a way that considers both men and women as mutually dependent entities (Chen, Fiske & Lee, 2009; Lee et al., 2010).

Consequently, the degree of satisfaction and conflict among women who were about to get married was measured by completing a questionnaire. In this questionnaire, they indicated the marriage myths they believed in, their premarital satisfaction, and their relationship confidence (Casad, Salazar & Macina, 2015). It was concluded that the degree of benevolent sexist attitudes is negatively related to the length of heterosexual relationships and their outcomes (Casad, Salazar & Macina, 2015; Leaper, Gutierrez & Farkas, 2022). This is explained by the conflict of heterosexual intimacy expectations between women who expect to depend on chivalrous men and men who expect to depend on caring, loving, and intimately related women (Chen, Fiske & Lee, 2009; Hammond & Overall, 2013; Leaper, Gutierrez & Farkas, 2022).

- Goal Achievement

Hammond and Overall (2015) conducted a study on heterosexual couples where participants were video recorded while discussing their independent personal goals with their partners. Couples also filled out questionnaires that assessed their goal-related competence, their relationship quality, intimacy, and their degree of endorsement of benevolent sexism using Glick and Fiske's Ambivalent Sexism Inventory. The results of their study indicate that women who endorse heterosexual intimacy are more likely to deviate their focus from their personal goals toward relationship-oriented support for their husband's needs (Hammond & Overall, 2015). On the other hand, men who endorsed heterosexual intimacy concentrate on providing dependency-oriented support which neglects their partner’s goals. Consequently, the presence of benevolent sexism in relationships impedes women’s competence and independent success while providing support for the fulfillment of men’s goals and intimacy needs (Hammond & Overall, 2015).

- Women’s Well-Being

Similar to the previous study conducted by Hammond and Overall (2015), heterosexual couples were asked to discuss a relationship problem that needed a change to be done from one of the partners’ sides. Several indicators were measured for the analysis, including the degree of endorsement of benevolent sexist attitudes and their correlation with women’s experiences during relationship conflicts (Cross, Overall & Hammond, 2016). It was concluded that women experienced heightened distress with their benevolent sexist partners, and this is due to feelings of insecurity and worthlessness in fulfilling their partners’ intimacy needs (Cross, Overall & Hammond, 2016).

To sum it up, even with the benign expression of benevolent sexism, its detrimental impact on women’s cognitive performance and physical health, the success/failure of heterosexual relationships, and gender inequality in society are inevitable. In addition, this seemingly positive expression of sexism is a double-edged sword that makes it difficult to identify, capture, and mitigate benevolent sexism. Thus, in the next section, we discuss previous work done in the field of sexism and hate speech detection in text.

2.2. Related Work in Machine Learning

In this section, we state previous studies' methodologies on the detection of sexism in its different forms, in addition to the research gap on benevolent sexism detection in text.

2.2.1. Hostile Sexism

Previous studies have mainly focused on the detection and identification of the hostile form of sexism through building a corpus of hostile sexist statements and training different machine learning models for the classification of these statements. For instance, Waseem and Hovy (2016) constructed a dataset to distinguish between sexist and racist tweets, and Waseem (2016) augmented the dataset with larger samples of sexist and racist tweets. They also added a new label "both" for statements that are both sexist and racist. The datasets were extracted using the Twitter Search API (Waseem & Hovy, 2016: Waseem, 2016). Specifically, terms and hashtags that are highly correlated with racism and sexism were searched, such as "WomenAgainstFeminism", "islam terrorism", and "gamergate" (Waseem & Hovy, 2016). In both papers, the Logistic Regression (LR) model was adjusted using different feature combinations to identify the optimal set of features (Waseem & Hovy, 2016: Waseem, 2016).

Moreover, the 16K tweet dataset constructed by Waseem and Hovy (2016) has been widely used as the basis for hate speech detection using traditional machine learning approaches like Support Vector Machines (SVM) and LR (Park & Fung, 2017; Frenda et al., 2019). In addition, deep learning neural networks and pre-trained language models like Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), FastText, and Bidirectional Encoder Representations from

Transformers (BERT) have been used in the classification of the hateful content in Waseem and Hovy (2016) dataset (Badjatiya et al., 2017; Park & Fung, 2017; Pitsilis, Ramampiaro & Langseth, 2018; Mozafari, Farahbakhsh & Crespi, 2020).

Further, Grosz and Conde-Cespedes (2020) focused their research on the context of workplace sexism by using previously available datasets (Waseem & Hovy, 2016: Waseem (2016): Goel, Madhok & Garg, 2018) and augmenting them with statements on workplace sexism from various online articles and other sources. In addition, four different versions of models (traditional and deep learning) were tested, and the optimal model was chosen for classification (Grosz & Conde-Cespedes, 2020).

2.2.2. Misogyny Detection

Similarly, there has been a wide research focus on the extreme form of hostile sexism, which includes misogyny and sexual harassment. Specifically, efforts have been directed toward creating a dataset for misogynistic speech classification based on specific hashtags and keywords on Twitter (Zeinart, Inie, & Derczynski, 2021: Anzovino, Fersini, & Rosso, 2018: Fersini, Rosso, & Anzovio, 2018: Fersini, Nozza, & Rosso, 2018).

Further, Frenda et al. (2019) study the differences and similarities between misogyny and sexism using the IberEval misogyny dataset (Fersini, Rosso & Anzovio, 2018), the Evalita misogyny dataset (Fersini, Nozza, & Rosso, 2018), and the sexist/racist dataset (Waseem & Hovy, 2016). Both the IberEval and the Evalita datasets have been extracted using the Twitter Search API using representative swear words and potential misogynistic Twitter accounts gamergate (Fersini, Rosso, & Anzovio, 2018: Fersini, Nozza, & Rosso, 2018). To illustrate, the study concludes that there is a high correlation between sexist and misogynist content (Frenda et al., 2019). This conclusion is based on the computational analysis which affirms that, in general, sexist statements hold hatred towards women, and in particular, misogyny (Frenda et al., 2019).

2.2.3. Multi-Label Categorization of Sexism

In addition, Parikh et al. (2019) attempted to create a neural framework to classify accounts of sexism in their dataset. They annotated the dataset collected from the “Everyday Sexism” website according to 23 labels of sexism. Further, their experiments on multi-label classification were done using traditional machine learning algorithms and deep learning neural networks. Further, they developed a neural framework that performed better than traditional algorithms and deep learning networks (Parikh et al., 2019). Further, Abburi et al. (2020) augment the previous work by creating a semi-supervised deep learning framework for the multi-label classification of sexism accounts.

2.2.4. Benevolent Sexism

More specifically, Jha and Mamidi (2017) constructed a corpus of both benevolent and hostile sexist tweets to be used in the detection and classification of sexism. The hostile sexist statements include sexist tweets extracted by Waseem and Hovy (2016). Further, the benevolent sexist tweets were extracted using the Twitter Search API and searching specific phrases “as good as a man”, hashtags “#adaywithoutwomen”, and patterns that have a high likelihood of being in the context of benevolent sexism (Jha & Mamidi, 2017). Jha and Mamidi (2017) used two models SVM and the Sequence-to-Sequence model for the classification of statements into benevolent, hostile, or others. SVM yielded higher results in comparison to the Sequence-to-Sequence model.

Intending to improve sexism detection and mitigation, Samory et al. (2021) augmented pre-existing datasets from Jha and Mamidi (2017) and Waseem and Hovy (2016) using the Twitter Search API with the “Call me sexist, but” search phrase, in addition to different psychological scales. Their dataset included the pre-existing statements under the categories “Benevolent” and “Hostile” and added statements under the categories “Sexism Scales”, “Call Me Sexist”, and “Other” (Samory et al.,2021). Further, they utilized traditional machine learning and deep learning models for the classification of sexist instances and concluded with the BERT deep learning model as the best performing among tested models.

As shown above, most of the datasets constructed focus on sexism in its harsh forms, the hostile and the misogynistic. The dataset of Jha and Mamidi (2017) was the first to include the Ambivalent Sexism Theory (Glick & Fiske, 1996) in the literature on sexism detection. However, since the data is relatively small and depends on the availability of tweets and retweets, more than 92% of the benevolent sexism data were lost after extraction and removing duplicates from retweets. Consequently, this lack of a reliable dataset is a major obstacle in tackling the issue of benevolent sexism, and it might lead researchers to limit their scope.

Thus, to enhance the detection of benevolent sexism, our research aims at creating an accurate classification system that detects this type of sexist statements. However, since the availability of pre-existing benevolent sexism datasets is unreliable, we are following a data-centric approach, which constitutes constructing a detailed and reliable corpus of benevolent sexism statements. This corpus will be collected from several sources and platforms to ensure diversity of expression. The methodology of collection and labeling will be clarified in the following section.

CHAPTER 3

DATA COLLECTION AND ANNOTATION

The process of creating a reliable dataset to be used in the detection of benevolent sexist statements comprised two main steps: data collection and data annotation.

3.1. Data Collection

Since benevolent sexism is present in a positive tone of expression, the search for such statements required looking into the occasions in which this type of sexism is manifested. Benevolent sexism was found in several forms including advice, quotes, and statements in informative articles. For instance, heterosexual intimacy was mostly expressed in husband-wife relationship quotes and advice. In addition, complementary gender differentiation was present in articles that set behavioral standards for women, and protective paternalism was found in quotes on how to treat, pamper, and protect women. Further, we included statements from articles that describe harmful social norms in societies, which were explained in a neutral manner that entailed potential benevolent sexist content.

To complement the potentially sexist content, we scraped empowering international women's day quotes. These statements were assumed to be among the non-sexist category and acted as an offset against benevolent sexist statements. Table 3.1 includes the titles of the articles used to extract data.

Also, we used the quotes' website "quotemaster.org" to scrape quotes based on specific search phrases and words such as "lady", "good wife", and "woman". These quotes made up almost 80% of the dataset.

Table 3. 1

Titles of Articles used to Extract Data

Article Title	Reference
Wife Quotes to Touch Her Heart	Sreekanth (2018)
How to Be the Woman Every Man Wants to Marry	Elhamy (2021)
How to be a Lady	Claytor (2022)
A brief history of Afghan women's rights	Gopalakrishnan (2021)
Ten harmful beliefs that perpetuate violence against women and girls	Veen, Cansfield, and Muir-Bouchard (2018)
28 Incredible and Empowering International Women's Day Quotes	Barrientos and Avendano (2022)
Share These 100 International Women's Day Quotes to Support Women's Rights	Liles (2022)

After analyzing the collected data, we created a list of patterns of statements that most likely include benevolent sexist attitudes. Then, we used Google Advanced Search to look up statements based on the patterns extracted such as:

- "you're" * "for a girl"
- a woman should be *
- "girls" * "as good as boys" *
- a good? wife should *

The final collection of data included 4,301 statements, which were partly annotated in the next section, and later used for the validation of the detection system.

3.2. Data Annotation

The pre-annotation process included detailed observation to identify the sexism categories present in the collected dataset. The statements included both hostile and benevolent expressions of sexism. Thus, we used the Ambivalent Sexism Theory (Glick

& Fiske, 1996) to annotate the statements into Hostile Sexist, Benevolent Sexist, or Non-Sexist.

The annotation was done by 9 graduate students at the American University of Beirut (6 Females and 3 Males), who have knowledge of gender studies and issues. A collection of 2,962 was distributed to all annotators, and the annotation was done in two phases (Figure 3.1):

3.2.1. General Categorization: Sexism vs None

- Sexism is the ideology that perceives one gender as prominent over the other.
- Non-Sexism is a neutral expression that has no affiliation towards any gender.

3.2.2. Specific Categorization: Hostile Sexism vs Benevolent Sexism

- Hostile Sexism: the sexism that's expressed in a negative, blatant, and aggressive manner, and it reflects men's hatred towards women.
- Benevolent Sexism: is the sexism that's expressed in a softer chivalrous tone, as it explains men's dominance through their care, need, and love for women.
- Table 3.2 shows examples of statements annotated according to the ambivalent sexism theory (Glick & Fiske, 1996).

The resulting dataset was comprised of 1,012 Benevolent Sexist statements, 733 Hostile Sexist Statements, and 1,217 Non-Sexist Statements.

Table 3. 2

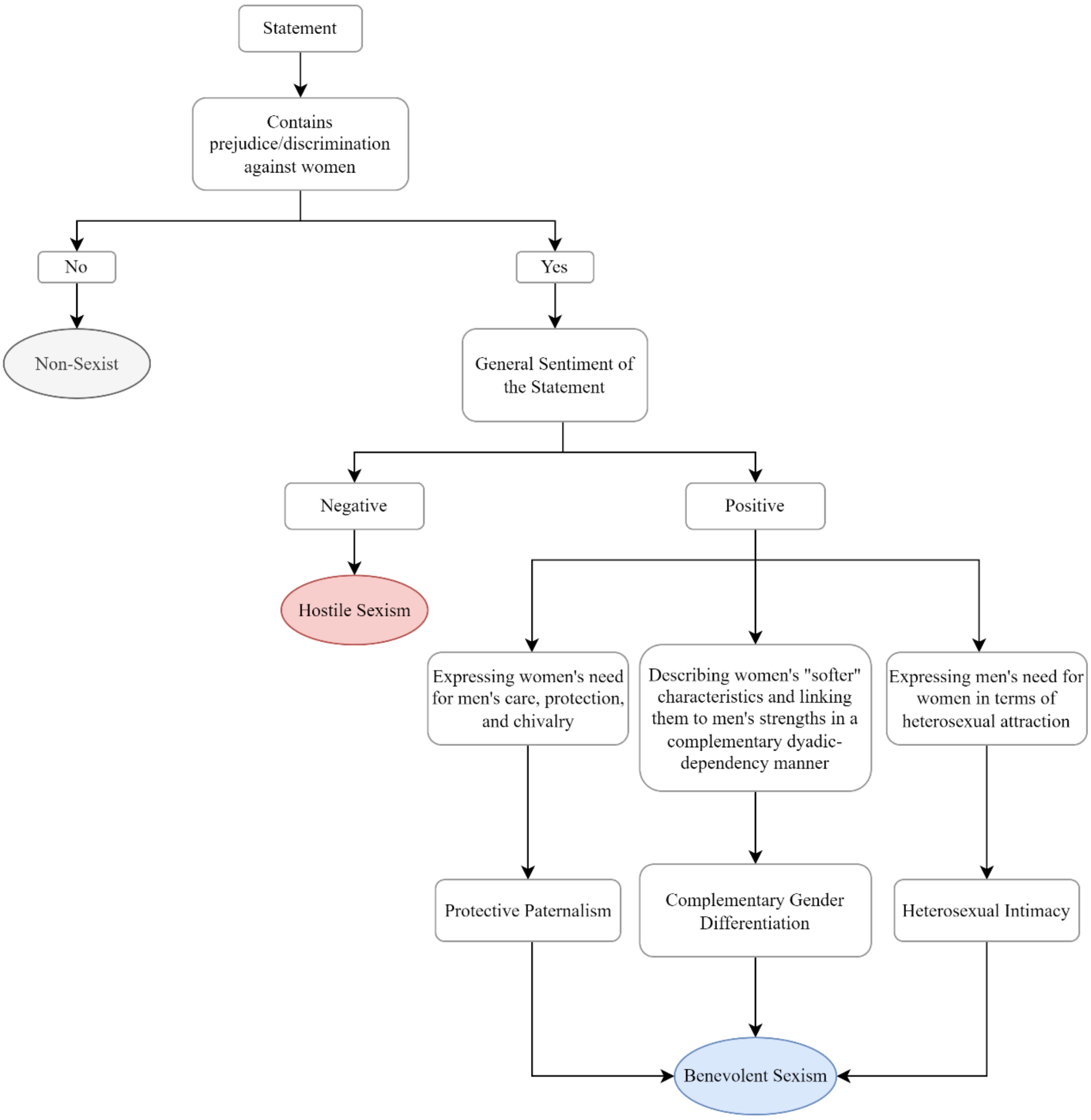
Sexism Examples

Statement	Annotation
Hold the door for a lady	Benevolent Sexism

The Devil is a woman	Hostile Sexism
There is no wrong way to be a woman	Non-Sexist

Figure 3. 1

Data Annotation Process


Statement
Contains prejudice/discrimination against women
No
Yes
Non-Sexist
General Sentiment of the Statement
Negative
Positive
Hostile Sexism
Expressing women's need for men's care, protection, and chivalry
Describing women's "softer" characteristics and linking them to men's strengths in a complementary dyadic-dependency manner
Expressing men's need for women in terms of heterosexual attraction
Protective Paternalism
Complementary Gender Differentiation
Heterosexual Intimacy
Benevolent Sexism

CHAPTER 4

EXPERIMENTAL METHODOLOGY

After creating a representative sexism dataset and achieving the first goal of our study, we moved on to utilizing this dataset in creating a detection system that allows for capturing benevolent sexist statements from large amounts of data. In this section, we elaborate on the experimental setup for training and testing several machine learning models using our collected and annotated data. The following is the process of choosing our best model to be used as the basis for a benevolent sexism detection system.

4.1. Data Preprocessing

The first step in our experiment was preprocessing and cleaning the collected data for machine learning. This included dropping duplicates and removing null values if any. In addition, we aim at creating a binary classification system, so we excluded the Hostile Sexist statements from our experimentation. Further, we constructed a function for normalization, which made the experimentation process more practical. This function included word tokenization followed by removing special characters, removing stop words, text lower casing, cleaning numbers, text lemmatization, and text stemming. Tokenization was first used to split the sentences into words. Then, lemmatization was used to switch a word into its base root mode (lemma) while taking into consideration the context of a sentence (Korenius et al.,2004). Further, the application of stemming reduced a word to its stem by stripping its suffix (Porter, 1980). The use of these information retrieval techniques was proven to yield better performance in comparison to baseline algorithms (Balakrishnan, 2014).

4.2. Feature Extraction

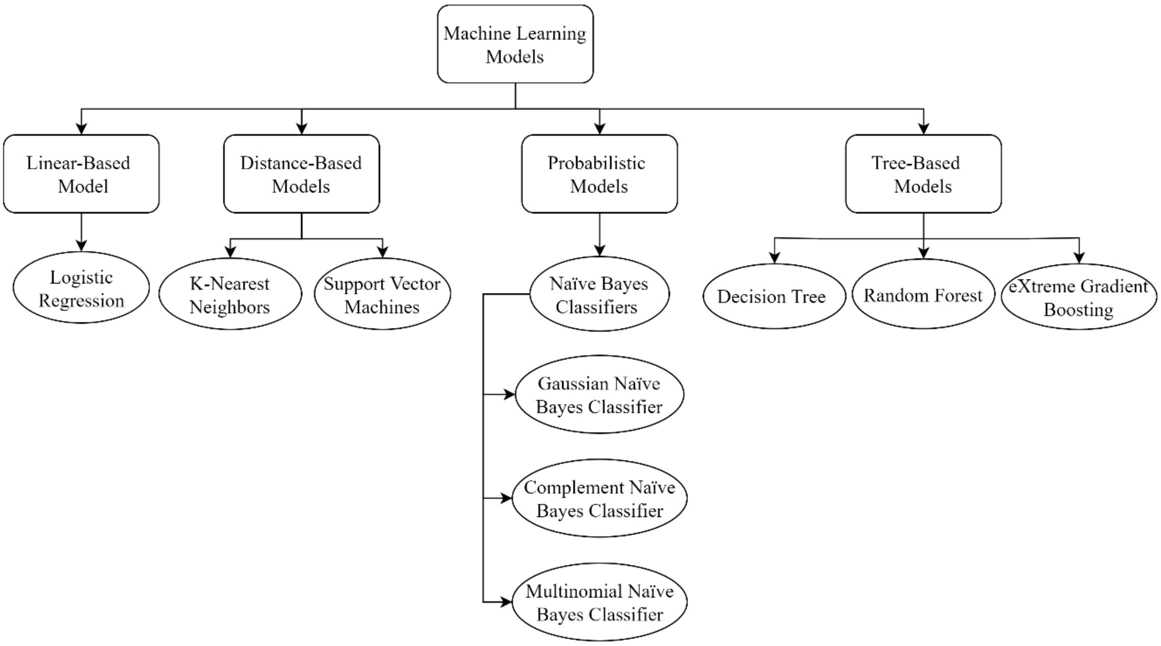
Since machine learning models cannot analyze textual data, we tested the models while comparing the results after two feature extraction techniques: CountVectorizer (CV) and Term Frequency-Inverse Document Frequency (TF-IDF). Using these tools, we were able to transform the textual data into numbers that can be analyzed by machine learning algorithms. In other words, the unstructured form of textual data was transformed into structured features.

The CV feature extraction tool transforms text into vectors according to the frequency of each word occurring in the text and thus, converts a set of strings into frequency representations (Harris, 1954). Moreover, TF-IDF also transforms a corpus of text into vectors of words/phrases counts while considering the importance of a word in a string in addition to its frequency in the entire set of strings (Luhn, 1957; Spärck Jones 1972). This means that the count of a word in a string can be counterbalanced by the number of strings that contain this word.

4.3. Machine Learning Models

Figure 4. 1

Machine Learning Models Used in this Study



As previously mentioned, we experimented using several machine learning classifiers (Figure 4.1) and tailored them according to our objective. The following are the models we used in our experiments:

4.3.1. Linear-Based Model

- Logistic Regression (LR): is a linear supervised machine learning algorithm for classification. It uses the logistic function to predict a dependent variable by analyzing the relationship between a set of independent variables (Hosmer & Lemeshow, 2013; Stoltzfus, 2011).

4.3.2. Distance-Based Models

- K-Nearest Neighbors (KNN): is a distance-based, supervised machine learning algorithm that can be used for both classification and regression. It is non-parametric, and it arranges data in a space that is defined by selected features (Fix & Hodges, 1989; Cover & Hart, 1967).
- Support Vector Machines (SVM): is a distance-based machine learning algorithm that distinguishes two classes of training data through an optimal classification function (Wu et al., 2007). This classification function is of a hyperplane that separates the data points with a maximum margin (Noble, 2006). Specifying this hyperplane maximizes the algorithm's correct classification of new examples (Noble, 2006). In our experiments, we use the Support Vector Classification (SVC) model provided by the Scikit-Learn library in Python.

4.3.3. Probabilistic Models (Naïve Bayes Classifiers)

Naïve Bayes Classifiers are based on applying the Bayes Theorem to classify data according to each point's feature vector (Rish, 2001). This application is associated with a "naïve" assumption that the presence of a feature in a class is independent of other features (Rish, 2001). Each of the following Naïve Bayes Classifiers follows a different statistical approach to computation.

- Gaussian Naïve Bayes (GNB): assumes a gaussian (normal) distribution of the class's probabilities (Horbonos, 2020).
- Multinomial Naïve Bayes (MNB): is a popular approach in natural language processing, as it considers average word counts in assigning feature vectors for classes (Sriram, 2021).

- Complement Naïve Bayes (CNB): calculates the probability of it belonging to the complement class/es and not to a particular class as in MNB (Rennie et al., 2003).

4.3.4. Tree-Based Models

- Decision Tree (DT): is a supervised machine learning approach that is used for both classification and regression. This model infers simple rules from the data features, concluding an if/else hierarchy of questions, that leads to a decision (Kotsiantis, 2013).
- Random Forests (RF): is implemented by randomly building a collection of decision trees. These decision trees might slightly differ from each other in the degree of overfitting and prediction performance. By averaging the results of the multiple decision trees, overfitting is decreased while the prediction power of the model is retained (Breiman, 2001).
- Extreme Gradient Boosting Ensemble (XGBoost): is a scalable and effective implementation of the Gradient Boosting Machines Framework (Friedman, 2001), which is another ensemble of decision trees that are combined to yield a more powerful model (Chen et al.,2015; Chen & Guestrin, 2016). What differentiates this model from RFs is the serial combination of the decision trees in a way that allows each model to correct the miscalculations of the previous one (Guido & Müller, 2016).

4.4. Performance Measurement

To reach the optimal benevolent sexism detection system, we used different classification metrics to measure the performance of the classifiers:

4.4.1. Confusion Matrix

The confusion matrix (Table 4.1) is a performance measurement tool for machine learning classifiers. It summarizes the number of correct and incorrect classifications, and it is essential in comparing the areas of error of a classifier.

Table 4. 1

Confusion Matrix

		Prediction	
		None-Sexist	Sexist
Actual	None-Sexist	True Negative (TN)	False Positive (FP)
	Sexist	False Negative (FN)	True Positive (TP)

4.4.2. Accuracy

The accuracy is the proportion of the correct classifications out of the total number of predicted samples.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

4.4.3. Sensitivity

The sensitivity measure, also known as Recall or True Positive Rate (TPR), is the percentage of correctly classified positive class statements out of all actual positive samples.

$$Sensitivity = \frac{TP}{TP + FN}$$

4.4.4. *False Positive Rate (FPR)*

The false positive rate, or 1-Specificity, is the percentage of incorrectly classified positive class statements out of all actual negative samples.

$$FPR = \frac{FP}{TN + FP}$$

4.4.5. *Precision*

The precision measure is the percentage of correctly classified positive classes of statements out of all positively classified samples.

$$Precision = \frac{TP}{TP + FP}$$

4.4.6. *F1 score*

By calculating the harmonic mean of a classifier's precision and recall, the F1 score combines both into a single metric.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

4.4.7. *Receiver Operating Characteristic Curve (ROC Curve)*

The ROC Curve (Figure 4.2) is a graph that displays the True Positive Rates and the False Positive Rates at different threshold settings. To clarify, the classifier predicts the probabilities of each statement belonging to the Sexist class and the probabilities of the same statement belonging to the Non-Sexist Class. Setting a threshold for the classifier indicates the minimum probability for a statement to be classified as Sexist. For instance, a threshold of 0.4 means that statements with predicted probabilities above 0.4 are classified as Sexist while statements with predicted probabilities below 0.4 are classified as Non-Sexist.

Further, different threshold settings yield different True Positive Rates and False Positive Rates, which are represented by the points in the ROC Curve.

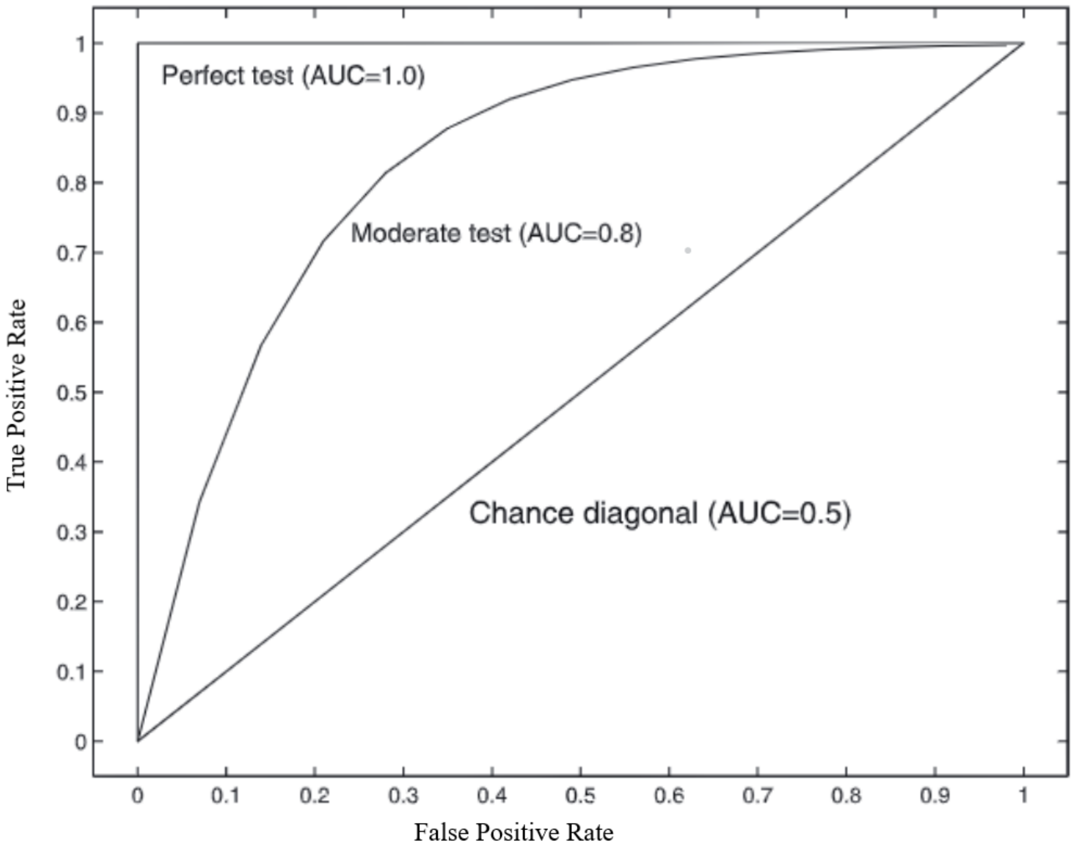
4.4.8. Area Under the ROC Curve (AUC)

The AUC score is an aggregate measurement of a classifier's performance across all potential classification thresholds. It indicates the classifiers' abilities to differentiate between the two classes (Sexist vs Non-Sexist) The perfect AUC score is 1.0, as represented in Figure 4.2.

Since we aim at capturing benevolent sexist statements, our focus was on true positives. However, since benevolent sexism is expressed in a positive manner, we were concerned about the model's ability to differentiate between sexist and non-sexist statements. Thus, we measured the accuracy in addition to the recall (sensitivity), precision, and FPR which were derived from the confusion matrix. Even though the confusion matrix yields an important set of classification metrics, this confusion matrix represents one operating point (Bradley, 1997). In other words, each classification threshold yields a different confusion matrix, and thus, different accuracy, recall, and precision. Consequently, we measured the overall performance of the classifiers across all classification thresholds with the AUC score.

Figure 4. 2

ROC Curve Explanation



4.5. Building the Models

Following excluding the hostile sexist statements and cleaning the dataset from duplicates, our final Benevolent vs Non-Sexist dataset consisted of 2,202 statements (55% Benevolent Sexist and 45% Non-Sexist). To avoid data leakage, we first split the data into 70% (1,541) training and a 30% (661) testing set.

Then, we used LR as a baseline model to test feature extraction and normalization techniques. For measuring the performance at this step, the model was run on a 10-Fold Cross Validation to validate the model's performance after each adjustment. During this application, the model was fit 10 times. With each iteration, the data was randomly split into 90% for training the model, and the remaining 10% as a hold-out for validation. The model generated the accuracy with each fitting, and the final cross-validation score was the average of the 10 accuracies.

We first ran the model on the dataset without normalization to choose the better-performing feature extraction technique. The CountVectorizer feature extraction tool resulted in higher classification performance in terms of cross-validation accuracy.

Table 4. 2

Feature Extraction Tools Cross-Validation Accuracies

Feature Extraction	Cross-Val Accuracy
TF-IDF	0.703402
Count Vectorizer	0.720943

Further, we iterated the model after applying different normalization techniques until the optimal combination is reached. This combination included word tokenization followed by text lemmatization and text stemming. This combination yielded a higher cross-validation accuracy.

Table 4. 3

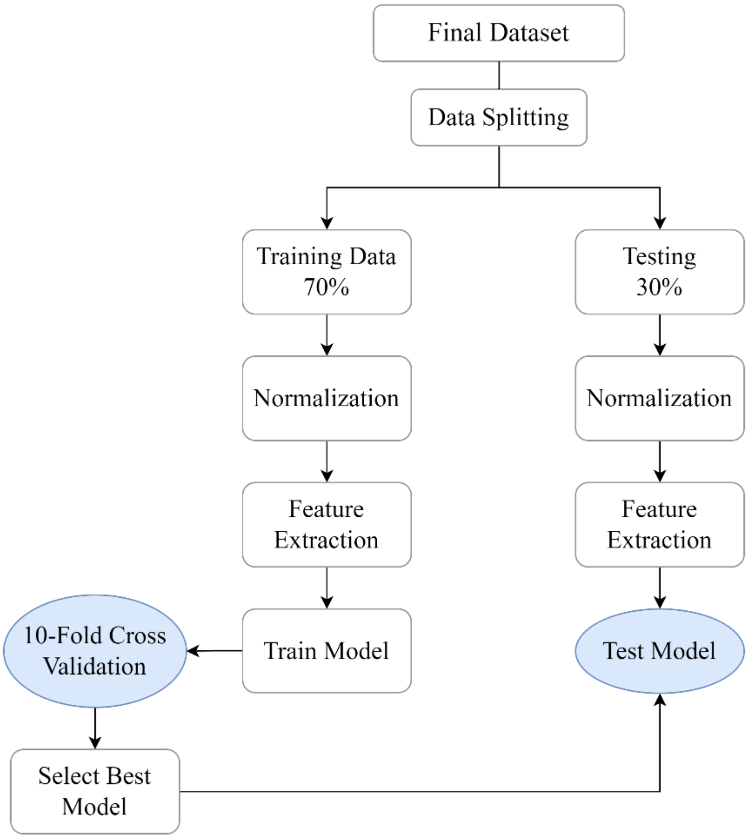
Preprocessing Tools Cross-Validation Accuracies

Preprocessing	Cross-Val Accuracy
Stemming Only	0.722225
Lemmatization Only	0.715086
Stemming & Lemmatization	0.726125

4.6. Choosing the optimal Model

Figure 4. 3

Models' Experimentation Steps



The next step was choosing the optimal model for the detection of benevolent sexism (Figure 4.3). During this step, we ran different models while measuring their classification performance. The cross-validation accuracy was measured. In addition, we tested the model on the test set to generate the AUC score of each classifier.

4.6.1. Results

Table 4.4 displays the cross-validation of each model using the train set. The XGB model showed the highest cross-validation accuracy (73.13%) followed by RF (72.29%), LR (72.09%), and SVC (71.09%). However, this metric is derived from the confusion matrix, which is in return constructed according to a specified threshold. Thus, each model might have automatically chosen an optimal threshold different from the other. Since this weakens the comparability between the models, we needed further measurements to reach the optimal classifier.

Table 4. 4

Models' Cross-Validation Accuracies

Model	Cross-Val Accuracy
eXtreme Gradient Boosting	0.731362
Random Forest	0.722878
Logistic Regression	0.720934
Support Vector Classification	0.71054
Decision Tree	0.683268
Multinomial Naive Bayes	0.672945
Complement Naive Bayes	0.671001
Gaussian Naive Bayes	0.654726
K-Nearest Neighbor	0.6431

We compared the AUC scores of the models and constructed the ROC-Curves on the test set for a visual comparison (Figures 4.4 and 4.5). As shown in Table 4.7, the SVC model had the highest AUC score (79.49%) followed by XGB (79.17%), RF (78.28%), and LR (77.39%).

Table 4. 5

Model's AUC Scores

Model	AUC Score
Support Vector Classification	0.794997
eXtreme Gradient Boosting	0.791694
Random Forest	0.782838
Logistic Regression	0.773874
Complement Naive Bayes	0.736254
Multinomial Naive Bayes	0.736254
Decision Tree	0.725094
Gaussian Naive Bayes	0.707404
K-Nearest Neighbor	0.700688

Figure 4. 4

ROC Curves of All Models

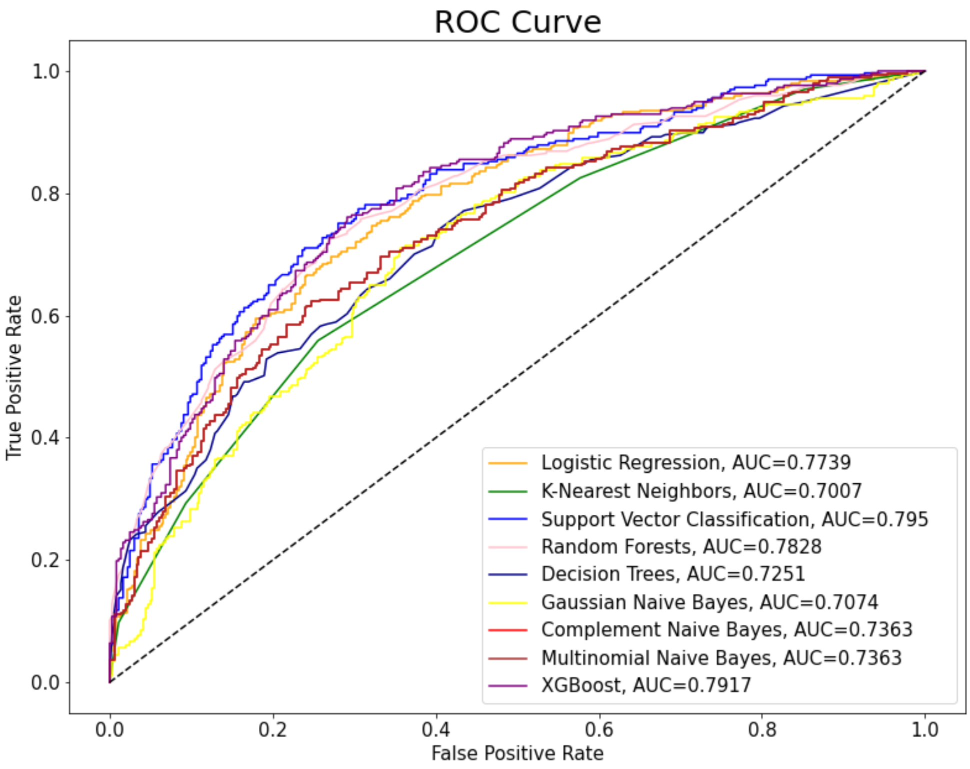
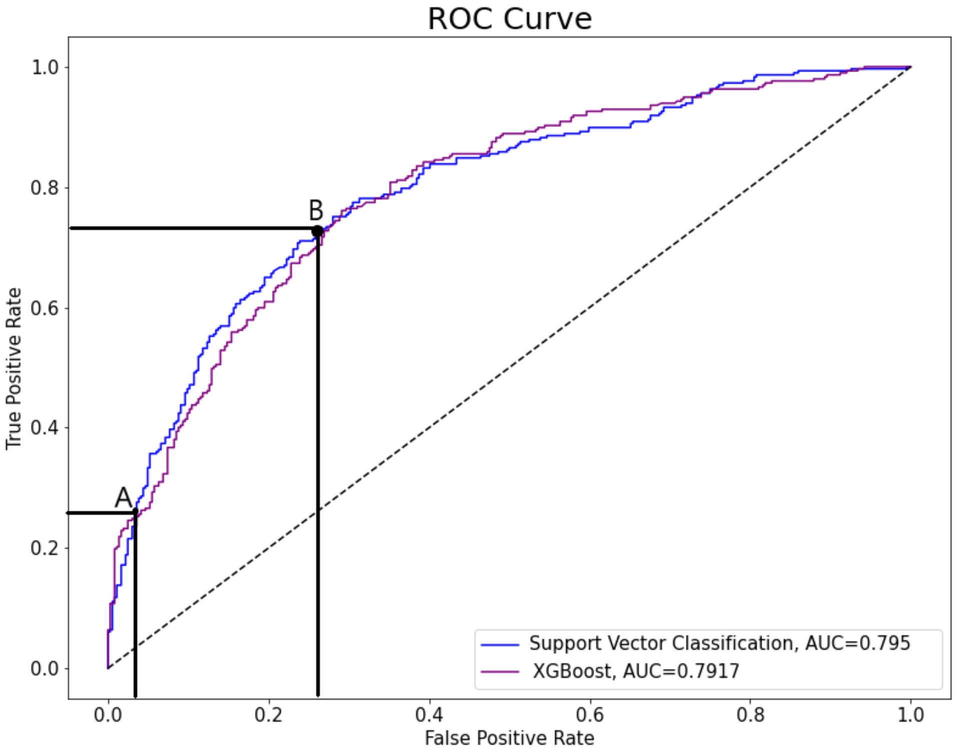


Figure 4. 5

ROC Curves for Best Models



4.6.2. Discussion

The main objective of our study is the detection of benevolent sexism. However, since both the benevolent sexist statements and the non-sexist statements might indicate benign and positive sentiments, we were concerned about the model's ability to differentiate between the two classes. Thus, we resorted to choosing the model with the highest AUC score, which efficiently complements the aim of the study. In other words, the AUC score proves the model's ability to distinguish between the two classes. A higher AUC score indicates better flexibility in threshold tuning, which can be associated with a significantly higher TPR and a tolerably higher FPR.

To illustrate, plotting all the models' ROC Curves (Figure 4.4) showed the leading models, which are SVC and XGB. Below point (A) and beyond point (B) in

Figure 4.5, XGB performs better in terms of a higher TPR for the same FPR. However, between points (A) and (B), SVC performs better in yielding a higher TPR for every FPR value. The [A, B] point range is optimal for defining threshold values since it includes TPR values between 26% and 73% and FPR values between 3% and 26%. Reconsidering our objective of efficiently detecting benevolent sexism, we consider an 15-18% FPR in the tolerable range while keeping the TPR above 60%.

Accordingly, the SVC model was observed as the best-performing classifier based on the AUC score and the ROC Curves. Even though the SVC model had a relatively lower cross-validation accuracy in comparison to the other models, we believed that hyperparameter tuning and threshold setting improves this metric (discussed in the next section). Also, the comparison of the AUC scores emphasized the SVC model's relatively higher ability to distinguish between the two classes (Benevolent Sexist vs None-Sexist) while keeping the FPR at a tolerable range.

4.7. Tuning the Model

After selecting the C-Support Vector Classification as the optimal model for our detection system, we moved to tune its parameters. This process involved identifying the target parameters, creating a list of potential values of these parameters, and running the model with different parameter combinations until the optimal one is reached. This process can be done manually by iterating the model using different parameters or using the GridSearchCV library function (Section 4.7.3), which is provided by the Scikit-Learn library in python.

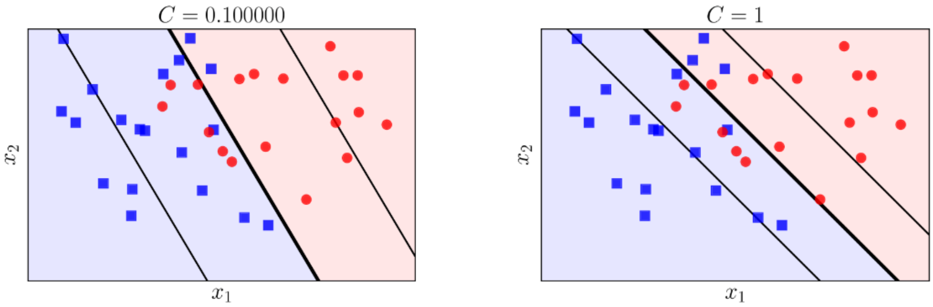
4.7.1. Defining Parameters to be Tuned

The parameter tuning was done on the three most critical parameters of the SVC (C-Support Vector Classification) model.: C-Regularization Parameter, Kernel, and the Gamma (Kernel coefficient).

- C-Regularization Parameter: As defined before, SVC is a distance-based machine learning algorithm that distinguishes two classes of training data through a hyperplane that separates the data points with a maximum margin (Wu et al., 2007; Noble, 2006). This imposed margin leads to possible misclassifications like the examples below (Dinh, 2019). The C parameter (Figure 4.6) gives the user control over misclassifications by increasing or decreasing the margin. The larger the C parameter, the smaller the margin, and the better performing the model is, and vice versa. However, a very large C leads to overfitting while a very low C leads to underfitting (Dobilas, 2021; Bzdok, Krzywinski & Altman, 2018). Thus, the goal is to find the optimal C, which can take any float value (“Support Vector Machines”, n.d.).

Figure 4. 6

C-Regularization Parameter



- Kernel: When the data cannot be linearly separable, a simple hyperplane as displayed before cannot solve the problem. In this case, the kernel parameter creates a multidimensional space that transforms the problem into a linear one and allows for solving the problem linearly (Figures 4.7 and 4.8) (Dobilas, 2021). The kernel can take values such as 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed', or 'callable' ("Support Vector Machines", n.d.).

Figure 4. 7

Kernel: Non-Linearly Separable Data

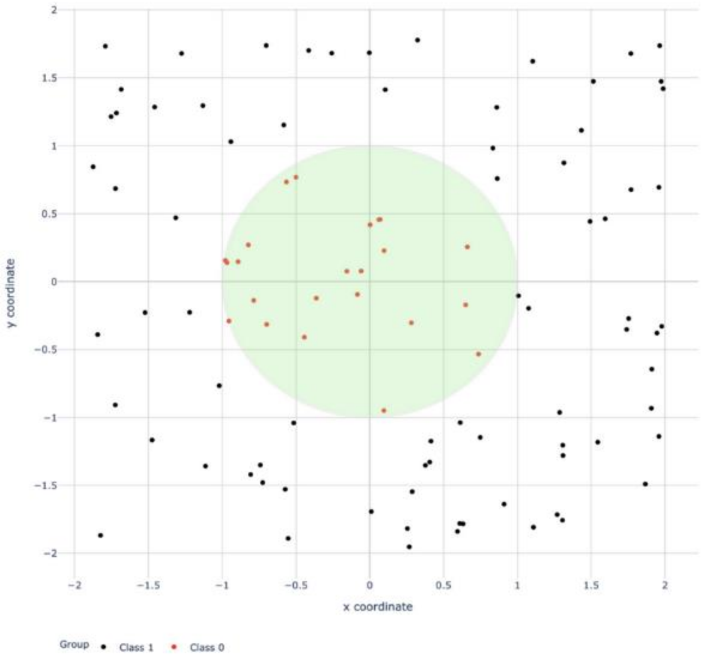
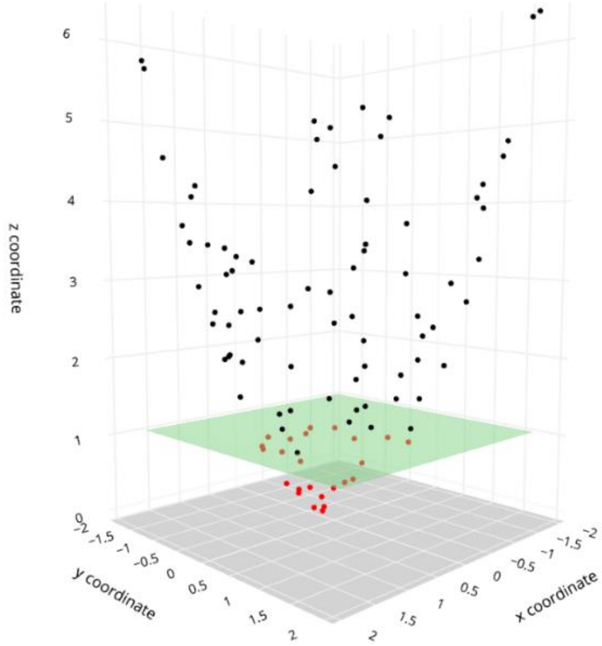


Figure 4. 8

Kernel: Multidimensional Space



- Gamma: When assigning the Linear Kernel, we need to adjust the C-Regularization parameter only. However, with Polynomial (poly), Gaussian Radial Basis Function (RBF), and Sigmoid kernels, the gamma is considered as a kernel coefficient that can be adjusted. The gamma determines the extent of points which affect the construction of the hyperplane ("Support Vector Machines", n.d.). High gamma (Figure 4.9) means that the radius of points affecting the hyperplane is small, and vice versa (Figure 4.10). Gamma can take values 'scale', 'auto', or a float number.

Figure 4. 9

High Gamma Illustration

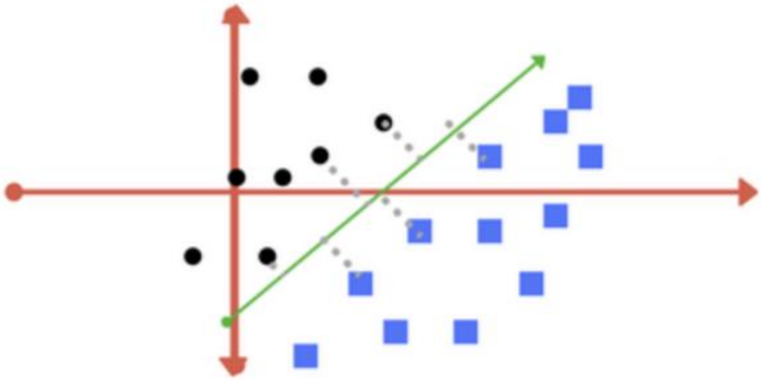
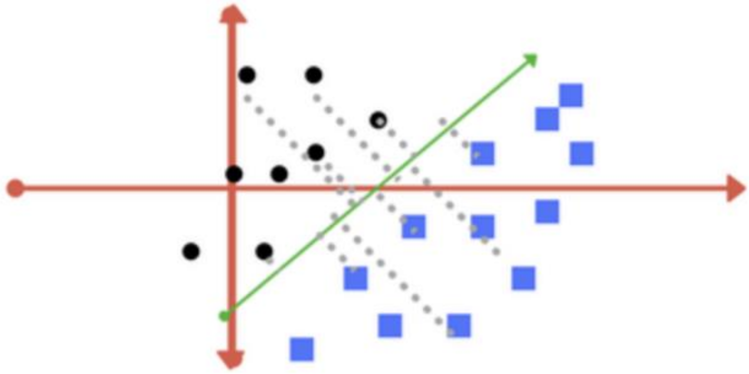


Figure 4. 10

Low Gamma Illustration



4.7.2. Manually Tuning the model

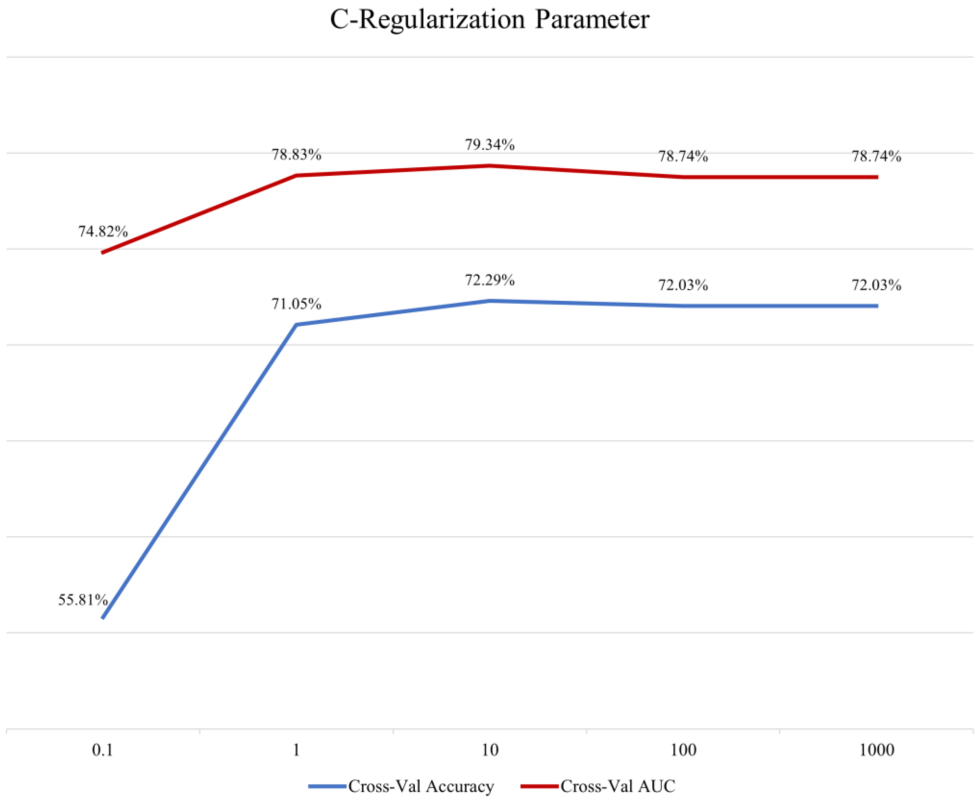
The model was manually tuned by changing one parameter at a time while keeping all other parameters at default. The default parameters for the C-Support Vector Classification model are C = 1, Kernel = ‘rbf’, and Gamma = ‘scale’. This step was done to narrow down the parameters to be tested using the GridSearchCV tool. For performance measurement, we applied a 10-fold cross-validation to the train set and we specified two classification metrics, which are ‘accuracy’ and ‘roc_auc’.

- C-Regularization Parameter

  The C-regularization parameter was experimented with the values [0.1, 1, 10, 100]. Figure 4.11 shows the classification metrics along different C values. The C = 10 yielded the highest cross-validation accuracy (72.29%) while C=1 yielded the highest cross validation AUC score of 78.83%. With C=100 and beyond, the performance of the classifier decreased and remained almost the same.

Figure 4. 11

Line Graph for Different C-Regularization Parameter Results

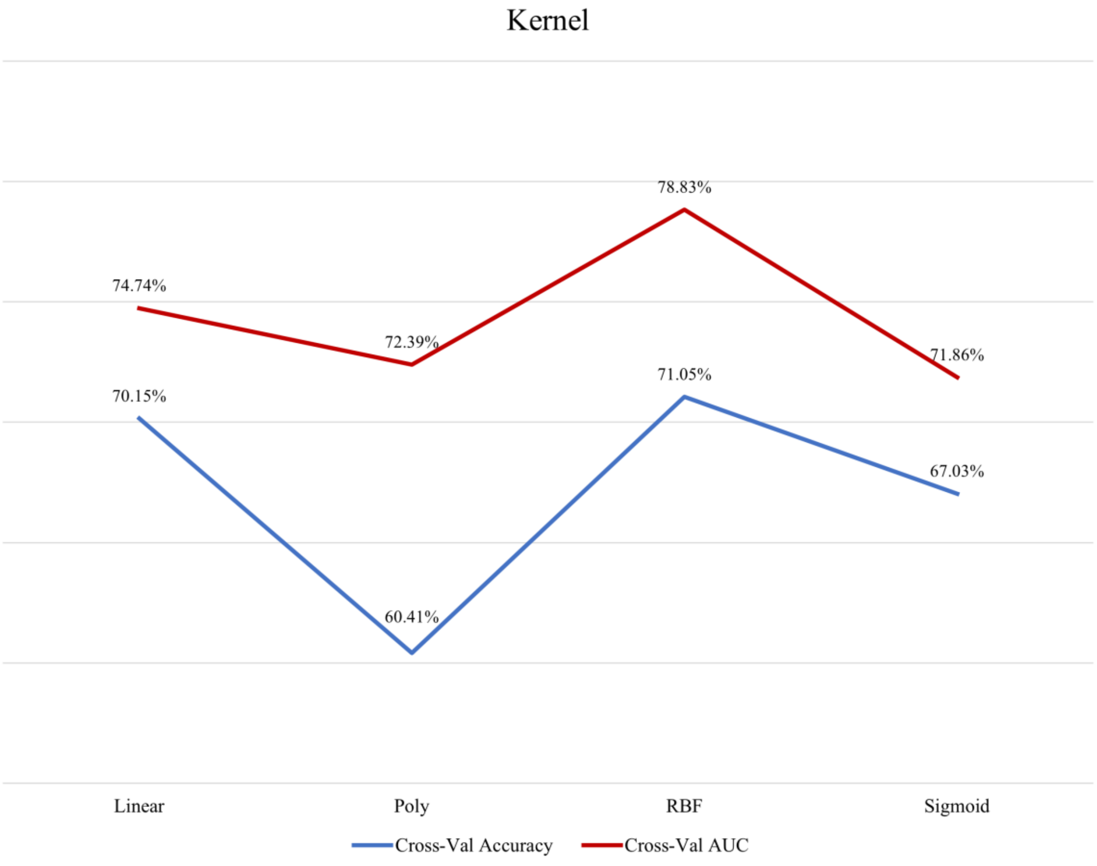


- Kernel

To test different kernels, we changed the values between [‘linear’, ‘poly’, ‘rbf’, and ‘sigmoid’] while recording the different results. As shown in Figure 4.12, the ‘rbf’ kernel showed the highest of all kernels, and thus, the best performance combined with the default parameters C = 1 and gamma = ‘scale’.

Figure 4. 12

Line Graph for Different Kernel Parameter Results



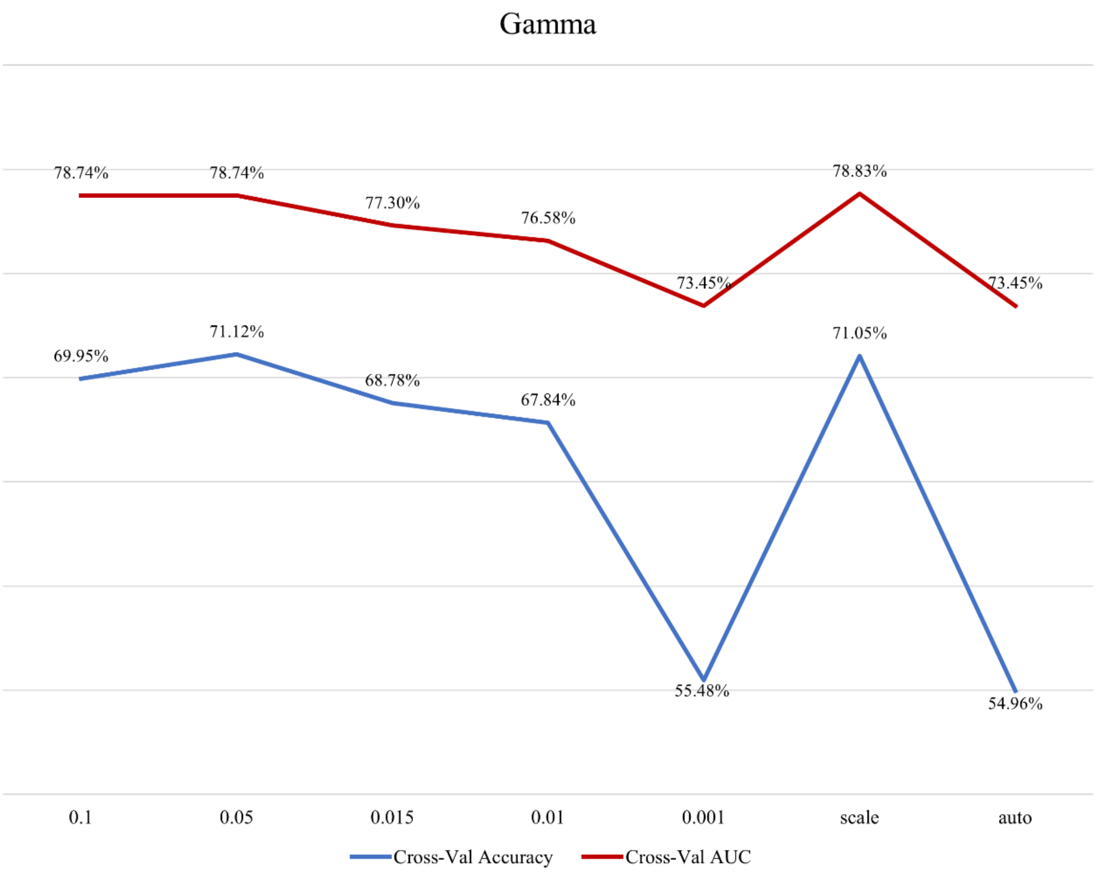
- Gamma

We tested the gamma with values [0.1, 0.05, 0.015, 0.01, 0.001, ‘scale’, ‘auto’] (Figure 4.13). Combined with the default parameters, C=1, and Kernel = ‘rbf’ (Gaussian

Radial Basis Function), the default gamma = ‘scale’ yielded the highest cross-validation accuracy (71.05%), and cross-validation AUC score (78.83%). In addition, gamma = 0.1 and gamma = 0.05 yielded relatively higher cross-validation accuracies (69.95%) and (71.12%), respectively, in addition to higher AUC scores (78.74%), respectively, relative to other gamma float values.

Figure 4. 13

Line Graph for Different Gamma Parameter Results



As shown above, the narrowed-down parameters to be tested in the following section are:

- C = [0.1, 10, 100]
- Gamma = ['scale',0.1, 0.05]
- Kernel = [‘rbf’]

4.7.3. Using GridSearchCV

The GridSearchCV is a model selection package provided by the Scikit-Learn library in python. After predefining hyperparameters of each classifier and specifying the scoring parameter (accuracy, F1 score, roc_auc, etc.), the function loops over all combinations of hyperparameters and fits the classifier on the training set in 5-Fold Cross Validation. With each iteration, the data is randomly split into 80% for training the model, and the remaining 20% as a hold-out for validation. The outcome of this step is the set of the model parameters that yielded the highest score.

As mentioned before, the model parameters experimented with were:

- C = [0.1, 10, 100]
- Gamma = ['scale',0.1, 0.05]
- Kernel = [‘rbf’]

With these parameters, the model had 9 parameter combinations and 45 fits using the 5-fold cross-validation. Since we are more concerned about the model’s high AUC score than a high accuracy, we specify the GridSearchCV scoring parameter to ‘roc_auc’. The function was run, and it concluded the following best estimator:

SVC (C = 10, gamma = ‘scale’, kernel = ‘rbf’, probability=True). The ‘probability = True’ parameter is used to extract predicted probabilities of each statement belonging to the benevolent sexist class and probabilities of the same statement belonging to the non-sexist class.

4.7.4. Evaluating the Chosen Model on the Test Set

Since the GridSearchCV function returned the model with the highest cross-validation AUC on the training set, we needed to compare the top 3 parameter combinations in terms of cross-validation AUC. Thus, we manually extracted the 5 AUC scores from the 5-fold cross-validation application, we calculated the standard deviation of each, and we chose the top three average AUC score combination. The C= = 10 performed best and was common in the three combinations. However, each gamma value yielded a different result with a different standard deviation (Table 4.6).

Table 4. 6

Cross-Validations and Standard Deviations of Top 3 Parameter Combinations

	Cross-Val AUC	Standard Deviation
C = 10, Gamma = 0.1, Kernel = 'rbf'	77.98%	1.186%
C = 10, Gamma = 0.05, Kernel = 'rbf'	78.20%	1.168%
C = 10, Gamma = 'scale’, Kernel = 'rbf'	78.22%	1.254%

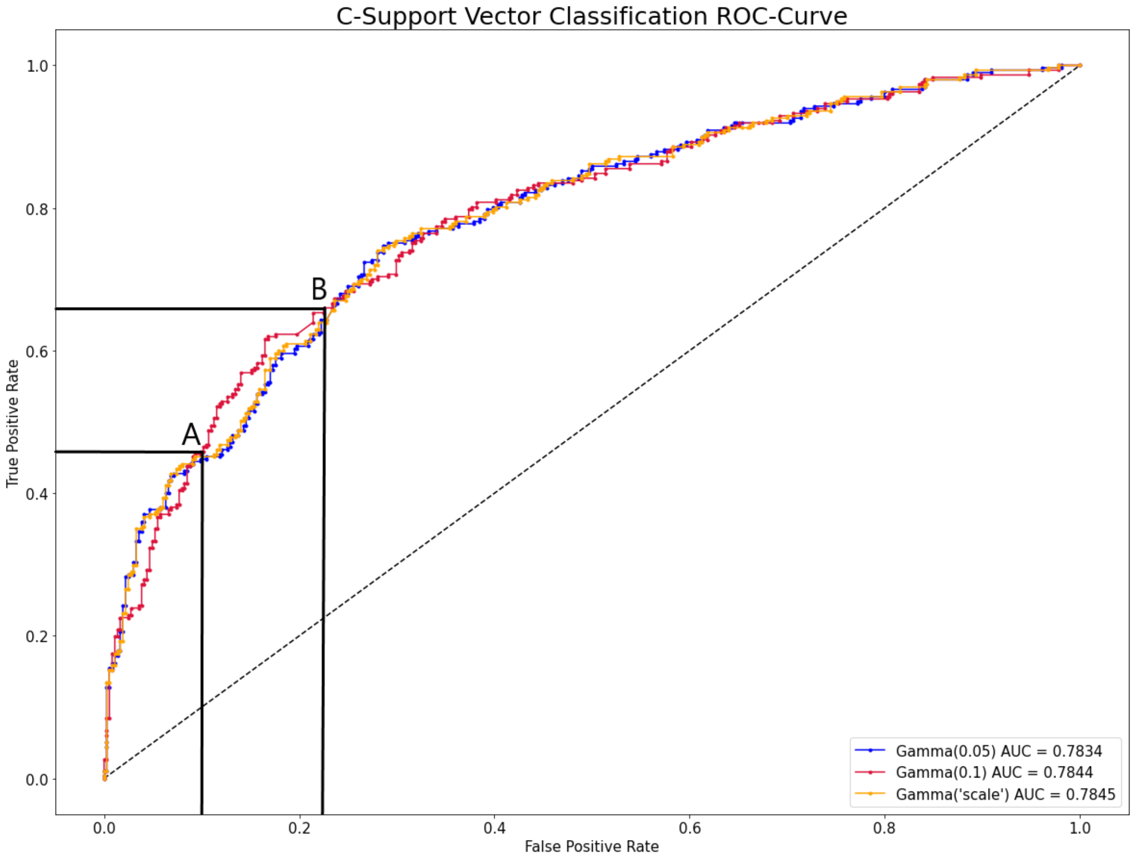
For this reason, we decided to apply the model on the test set and generate the respective ROC Curve for each parameter combination (Figure 4.14). As shown below, gamma of 0.05 and ‘scale’ values yield very similar ROC Curve shapes. However, the gamma = 0.05 gives the highest test AUC score followed by gamma = 0.1 (78.45% and 78.44% respectively).

As discussed in Section 4.6.2, a higher AUC score indicates better flexibility in threshold tuning, which can be associated with a significantly higher TPR and a tolerably higher FPR. However, since the two parameters, gamma = 0.05 and gamma = 0.1, close test AUC score, we resorted to choosing the range of desired TPR and FPR values as done previously.

To illustrate, plotting all the models' ROC Curves (Figure 4.14) showed the leading models. Below point (A) and beyond point (B) in Figure 4.14, gamma values of 0.05 and 'scale' yielded better performance in terms of a higher TPR for the same FPR values. However, between points (A) and (B), gamma=0.1 leads to a better performance in yielding a higher TPR for every FPR value. The [A, B] point range is optimal for defining threshold values since it includes TPR values between 46% and 64% and FPR values between 10% and 23%. As previously stated in Section 4.6.2, we consider a 15-18% FPR in the tolerable range while keeping the TPR above 60%. As a result, we choose the SVC model with the parameter combination of C = 10, Gamma = 0.1, and Kernel = 'rbf'.

Figure 4. 14

ROC Curves of Top 3 Parameter Combinations on SVC Model



4.8. Threshold Tuning

The last step in choosing our optimal model was to choose the optimal classification threshold according to our objective. As previously mentioned, the two classes under study are of neutral-positive expression. Consequently, there was difficulty in distinguishing benevolent sexist statements from non-sexist ones. For this case, we were concerned about the model's ability to maximize the TPR while keeping the FPR at a tolerable level.

The following analysis was fully based on the test set. We experimented with several thresholds until the set present in Table 4.7 was reached. These thresholds were chosen according to the ROC Curves. As seen in Figures 4.15 and 4.16, the selected thresholds represent outward stiff points on the ROC Curve. In other words, these specific points meet our criteria for selecting thresholds that yield a TPR of higher than 60%, a tolerable FPR, and all while in the range where this SVC model outperformed other versions with different parameter combinations (Section 4.7.4).

The highest threshold of 0.5487 means that all statements with predicted probabilities above 0.5487 were considered benevolent sexist statements. This threshold yielded a 56.90% TPR and 14.01% FPR. To increase the recall (TPR), the increase in the FPR was inevitable. Thus, after further decreasing the thresholds, we meet the optimal value of 0.5185, which yields a TPR higher than 60% (61.95%) and an FPR of 16.76%.

Table 4. 7

FPR and TPR of Different Threshold Settings

	Threshold			
	0.4919	0.5185	0.53	0.5487
True Positive Rate	65.32%	61.95%	59.26%	56.90%
False Positive Rate	21.43%	16.76%	16.21%	14.01%

Figure 4. 15

ROC Curve with Plotted Threshold Settings (1)

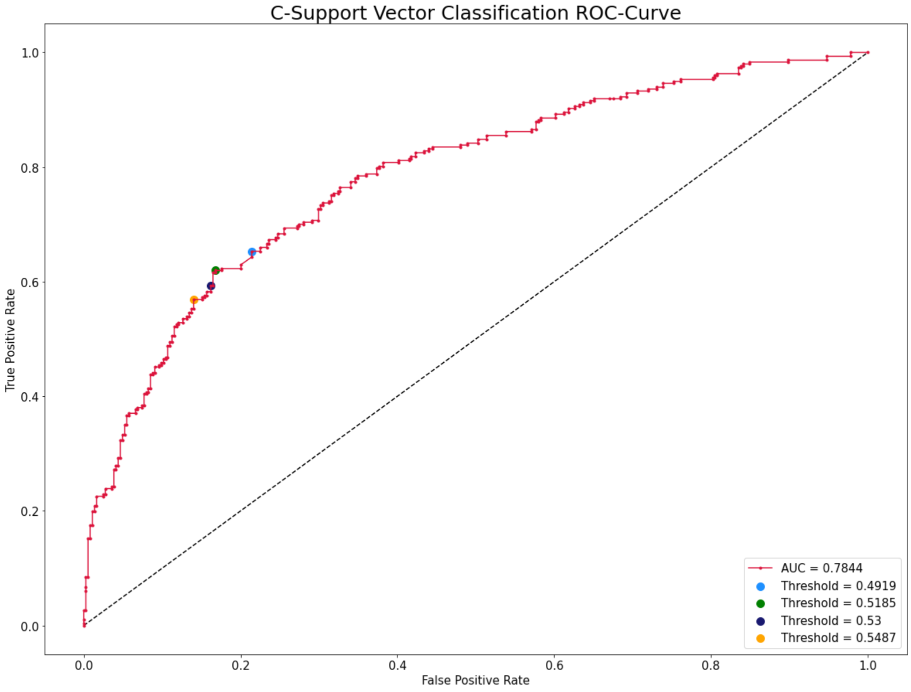
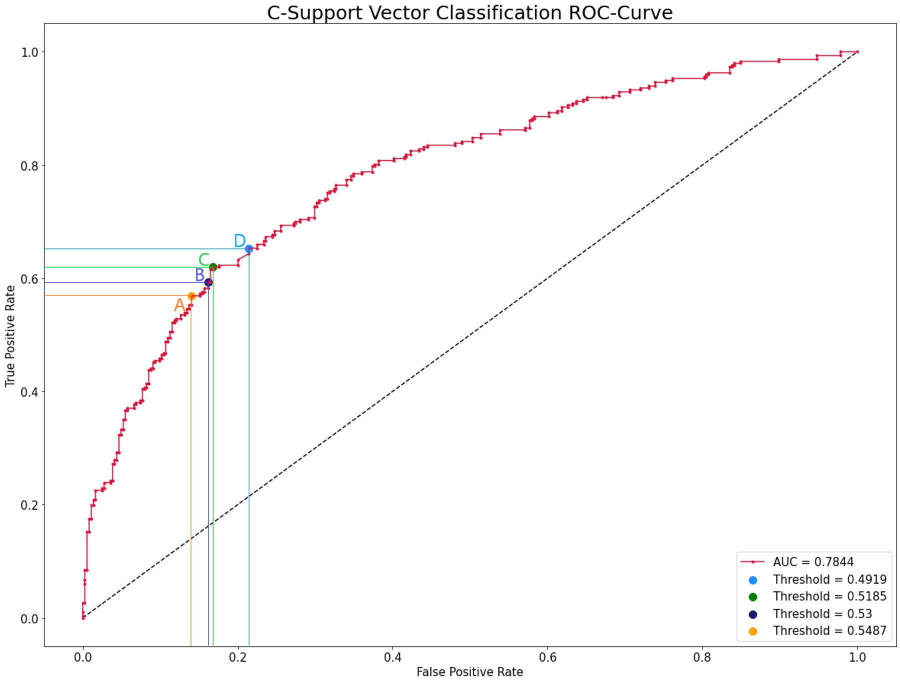


Figure 4. 16

ROC Curve with Plotted Threshold Settings (2)



The final step of testing our tuned model is returning the classification performance measures for comparison in later stages. Table 4.8 presents the classification metrics of the final tuned model: SVC with parameters C = 10, Gamma = 0.1, and Kernel = ‘rbf’. This model has a threshold set to 0.5185, meaning that all statements with predicted probabilities above 0.5185 were considered benevolent sexist statements.

Table 4. 8

Classification Performance of the Tuned Model

TUNED SVC	
Accuracy	73.22%
Precision	74.19%
Recall	61.95%
F1-Score	67.52%
AUC Score	78.44%

Finally, we saved our model in a pipeline that included the feature extraction step using CountVectorizer and the SVC(C=10, gamma=0.1, kernel=’rbf’,). Then, we fit the model on the training set, and we saved it for the validation step.

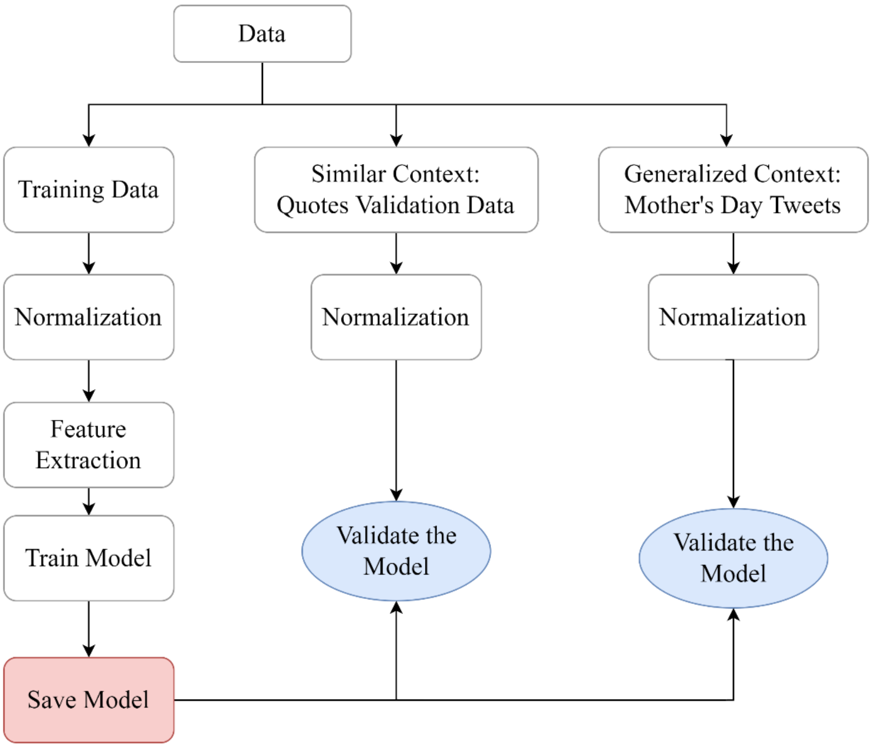
CHAPTER 5

MODEL'S GENERALIZABILITY: CASE STUDY

After constructing and tuning the classification model, we moved to validate the classification performance of the model on unseen data in two approaches. The first approach to our case study constituted using the model to classify the remaining unlabeled quotes from the collected data. This step measured the effectiveness of the model on a similar context dataset. The second approach to our case study measured the effectiveness of the model in a generalized context. This step allowed us to capture the strengths and weaknesses of the model for further fine-tuning and adjustments.

Figure 5. 1

Model Validation Steps



5.1. Similar Context Quotes

Following the data annotation, a remaining part of the dataset (1,339 statements) was left unlabeled, and the machine learning models were not trained nor tested on it. Thus, it was used as a verification means for the model's success. Initially, our annotated dataset was split into 70% training and 30% testing. This meant that the model was trained on 1,541 statements, tested on 661 statements, and would be validated on 1,339 statements. We manually labeled the validation dataset before the model's application, and it contained 260 (19.42%) benevolent sexist statements and 1,079 (80.58%) non-sexist statements.

Further, the validation dataset was normalized using the previously applied normalization methods: tokenization followed by lemmatization then stemming. Further, we imported the dumped pipeline which included the feature extraction tool and the model. The statements probabilities were predicted and the previously selected threshold of 0.5185 was set when predicting class labels.

5.2. Mother's Day Tweets

The second part of our validation process included a broader scope for the model's application: Mother's Day on Twitter. We chose this occasion since it contains various forms of expression on women's contributions to their households and a potential presence of benevolent sexism in such expressions. On March 21st, 2022 (Mother's Day), we used the Twitter API to extract tweets while specifying the search keywords to 'a mother's love', 'a mother is', 'every mother and wife', and 'a good wife'. A collection of 4,267 tweets were collected. We manually labeled 971 tweets to be used for the validation of the model. This collection included 95 (9.8%) benevolent sexist statements and 876 (90.2%) non-sexist statements.

The same process was applied to this validation dataset, which included tokenization, lemmatization, and stemming. Then, we used the pipeline to predict the class labels and the probabilities of the statements belonging to the benevolent sexist and non-sexist classes.

5.3. Results

To measure the classification performance on the validation sets, we returned the recall (TPR), and FPR, which are a result of specifying the model's threshold to 0.5185, and the AUC score of each iteration. Since both datasets were imbalanced, the accuracy score could be misleading. Thus, we calculated the precision score as a representation of the model's accuracy in detecting the minority class (Benevolent Sexism).

Table 5. 1

Classification Performance of the Model on the Test Set and Validation Datasets

	Testing	Validation (Quotes)	Validation (Tweets)
Recall (TPR)	61.95%	46.79%	81.05%
False-Positive Rate (FPR)	16.76%	13.59%	10.17%
Precision	74.19%	45.93%	46.39%
AUC Score	78.44%	72.89%	91.48%

To gain a clearer look into the model's performance on both unseen datasets, we draw the respective ROC Curve for each application (Figures 5.2 and 5.3). The ROC

Curves allowed us to visualize the TPR and FPR trends over different threshold settings.

Figure 5. 2

Validation on Quotes ROC Curve

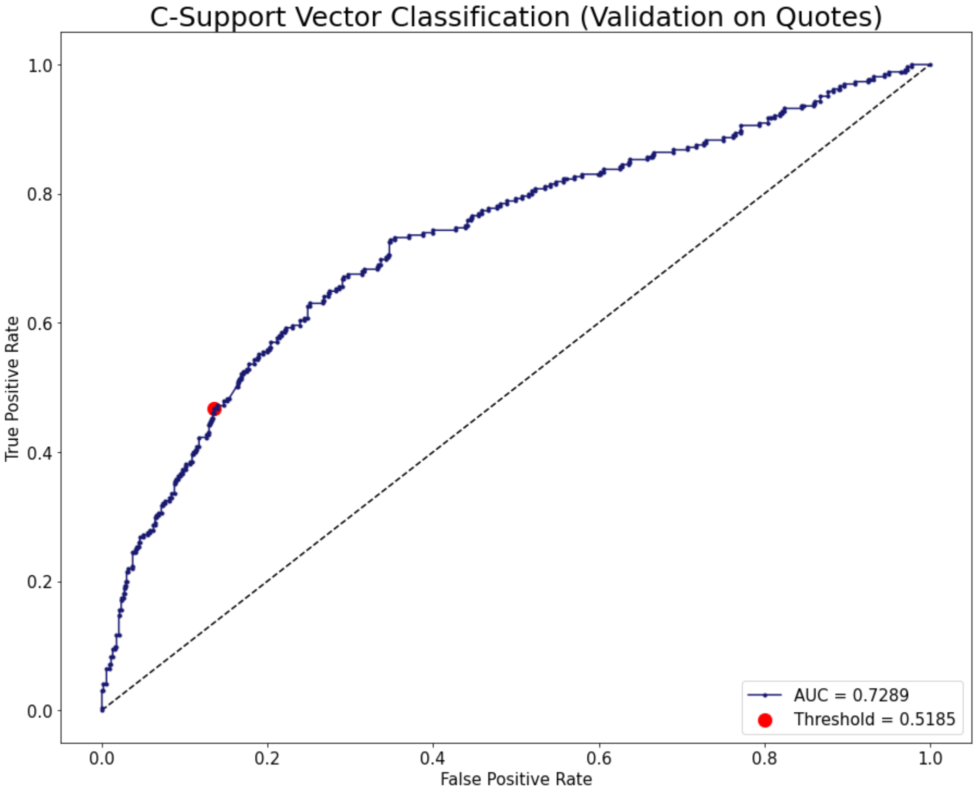
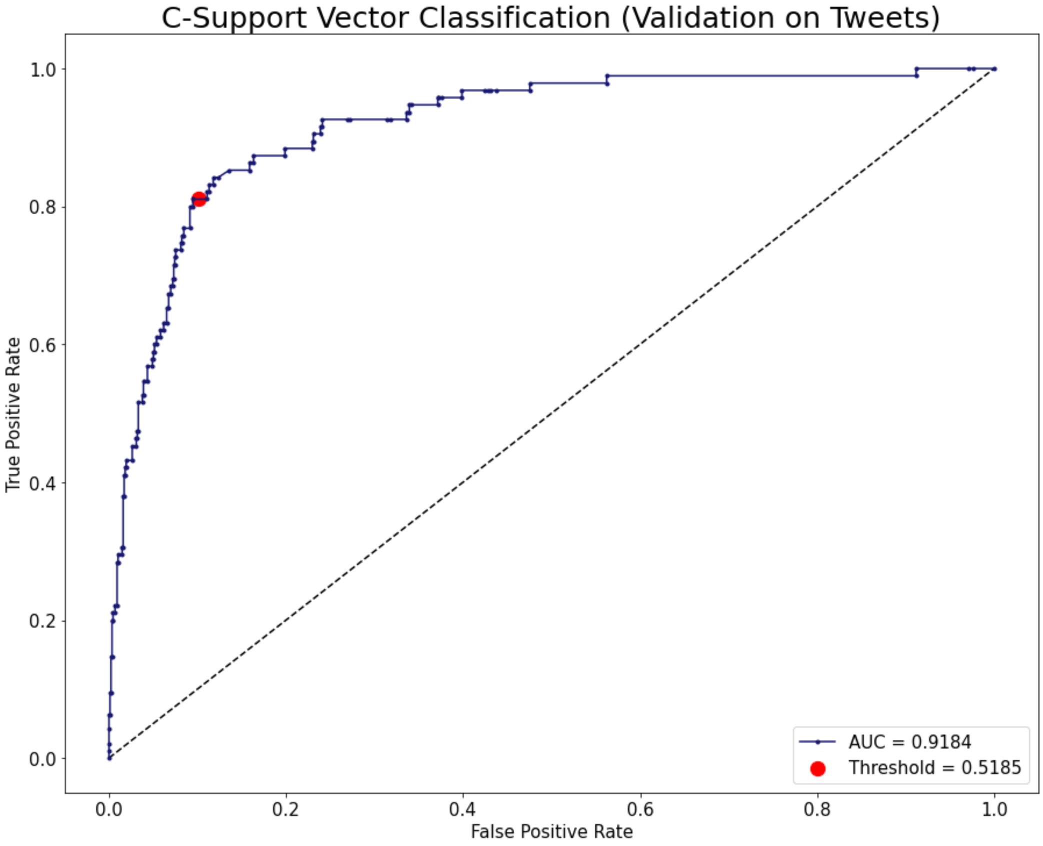


Figure 5. 3

Validation on Tweets ROC Curve



Overall, the AUC scores for both validation experiments presented a significant difference in performance: a 7.08% decrease in the quote's validation dataset and a 16.6% increase on the Mother's Day tweets dataset. These results showed a similar model performance on distinguishing the positive class (benevolent sexist) and the negative class (non-sexist) in the unseen quotes dataset, but a significantly higher model performance on the Mother's Day tweets dataset.

Further, the model's TPR and the FPR on the testing set were 61.95% and 16.76% respectively, which are based on the set threshold of 0.5185. On the classification of the quote's validation dataset, the model's TPR and FPR were 46.79% and 13.59%, respectively. The latter could be viewed in Figure 5.2.

On the other hand, the model's TPR and the FPR on the classification of Mother's Day tweets were 81.05% and 10.17%, respectively. When plotting the coordinates in Figure 5.3, we can observe the set threshold of 0.5185 that yielded these results.

In both applications, the precision score significantly decreases, showing that the accuracy of the minority class (Benevolent Sexism) is low. To go further into the details of errors in the model classification, we performed an error analysis in the next section.

5.4. Error Analysis Discussion

The model had a lower precision score in both applications on unseen datasets. Thus, to find the areas of error we performed an extensive error analysis of the model's misclassifications.

5.4.1. False Positives

Statements that were incorrectly classified as benevolent sexist belonged to several categories.

- Husband/Wife: We observe a high number of false positive statements where husbands and wives are stated in a comparison or connection manner. Almost 24% of false positive quotes and 18% of false positive tweets included both words 'husband' and 'wife', and 66% of the false positive quotes included the word 'husband'. Table 5.2 presents some examples of statements misclassified as benevolent sexist under this category. Such statements and several others were considered benevolent sexist because they express an issue in the marriage context. Also, statements comparing or connecting men to women led the model to classify them as benevolent sexist even when the statement advocates gender equality. This might be because the marriage context occurred continuously in the training set in statements that refer to the husband and wife in a sexist manner, such as “the true lady is in theory either a virgin or a lawful wife”.

Table 5. 2

False Positive Examples (1)

	Statement
1	The relationship between husband and wife should be one of closest friends.
2	For years, my husband and I have advocated separate vacations. But the kids keep finding us.
3	LOL Pastor is a former US Air Force vet and Republican. Haven't met his wife but he's real nice. Just social w ever…
4	This is definitely how you feel. You don't have to be a wife or a husband to be a good person or not to be coping ….

- Woman: The word woman occurred continuously in false positive classified tweets and quotes (Table 5.3). The word woman appeared in 38% of the false positive quotes and 52% of the false positive tweets. We observe a large number of misclassified statements containing this word since it has been previously associated with sexist context such as “A married woman, the safest place for any woman to be is at home with her husband” and “A woman without perfume is a woman without a future". This can be highly associated with the fact that 50% of the statements in the training set contained the word "woman".

Table 5. 3

False Positive Examples (2)

	Statement
1	Christ was born of a woman without the man.
2	I'm a Self-made Woman in Every Sense of the Word
3	SALEM, Mass. — A murder trial is scheduled next month for a Haverhill man accused of killing a woman outside a Lawr…
4	Post-pandemic, one in every three health insurance policies sold was to a woman, a study conducted by SBI shows.…

- Transgenders in sports debate: During the time of the year when the tweets dataset was extracted (March 21st, 2022), the debate on transgender women participating in women's sports had just elevated, so a lot of people were tweeting on the topic while mentioning male and female physical characteristics in sentences. This led to almost 13% of the false positive tweets (Table 5.4).

Table 5. 4

False Positive Examples (3)

	Statement
1	Ths person is not physically a woman! Dot u get it???
2	@ErikaDonalds this is unfair. A woman is biological not build as a man. these are just facts.
3	He is not a woman, he is a human being full of testosterone; women's sport is over!

- Sexist Patterns: The model was trained on sexist statements such as “Think like a lady. Act like a man” and “a good wife makes a good husband”. Consequently, once the mode captured these patterns in the quotes and tweets, it classified them as benevolent sexist. Table 5.5 shows some examples of misclassified statements due to the presence of benevolent sexist patterns in them.

Table 5. 5

False Positive Examples (4)

	Statement
1	Women like me because I don't look like a girl who would steal a husband. At least not for long.
2	Every woman is the architect of her own fortune.
3	Dylan was again writing some of the best love songs in the genre, like "Visions of Johanna," "Just Like a Woman," and "Sad-Eyed Lady of the Lowlands."

5.4.2. False Negatives

Further, looking into the false negatives, we saw that the model performed well in identifying none-sexist statements as none. However, taking a closer look into the false negatives in the unseen quotes dataset, we can see that almost 38% of the falsely classified are long statements (containing 30 words and more). However, looking back to the false positives, only 4% were long statements. The model’s performance seemed to deteriorate on long sentences and classify them as non-sexist. Table 5.6 displays some examples of the false negatives from the quotes dataset.

Table 5. 6

False Negative Examples (1)

	Statement
1	My own grandmother went to great lengths to make sure I knew simple things like how and when to open the door for a lady. And the best thing my mama taught me was to pray.
2	And being a husband made me helpless, because I had somebody to protect (somebody a little high-strung, who had a tough time emotionally with things like the lights going out indefinitely).
3	No young lady can be justified in falling in love before the gentleman's love is declared, it must be very improper that a young lady should dream of a gentleman before the gentleman is first known to have dreamt of her.

On the other hand, the model’s performance on the tweets dataset yielded significantly better results in classifying non-sexist tweets. However, some tweets were misclassified as non-sexist. As shown in Table 5.7, benevolent sexism is embedded in the meaning of the tweet. For instance, tweet 1 manifests notions of benevolent sexism through expressing the characteristics of a perfect wife as “classy beautiful”, in addition to her responsibility of taking good care of her husband.

Table 5. 7

False Negative Examples (2)

	Statement
1	@The_3_God I vote Iyah she is classy beautiful She takes very good care of you. She is the epitome of a wife. She…
2	@clariss55827655 He used to brag that he made good money. Enough money that his wife could stay home. I don't know…
3	@idfkkkk_____ @uwantaqua Not reply jellly. If anything I want a wife. A black female. So what’s good. She’s my baby…

5.5. Discussion and Recommendations

Following the construction and tuning of our SVC model for the classification of benevolent sexism in text, we noticed a disparity in the model's performance on the two validation datasets.

When validating the model on the unseen quotes dataset, it showed a consistent performance in terms of the AUC and FPR scores but lower performance in terms of TPR. However, in the broader context application (tweets dataset), the model performed better in terms of the TPR, FPR, and AUC scores. In both applications, the model yielded a lower precision score.

Since the TPR, FPR, and precision scores are highly correlated with the threshold setting, the analysis of the Validation on Quotes ROC Curve present in Figure 5.2 indicated our capability of increasing the TPR by decreasing the previously set threshold.

However, the Validation on Tweets ROC Curve present in Figure 5.3 showed exceptionally high performance in terms of the model's generalizability with an AUC score above 90%. The model revealed a high ability of distinguishing benevolent sexist from non-sexist statements despite the positive connotation that was associated with Mother's Day tweets. This performance, despite the high data imbalance, indicated our accomplishment in creating a representative and generalizable benevolent sexism dataset.

Following the error analysis, we concluded that the model was misclassifying statements that referred to certain contexts. For instance, tweets and quotes containing the words "lady", "husband", and "wife" were misclassified as benevolent sexist statements. Also, tweets relating to the debate on transgender women in sports were misclassified as benevolent sexist since they include a comparison of men’s and women’s physical characteristics. Further, the model misclassified long quotes (30 and more words) as non-sexist, which revealed a weakness in detecting benevolent sexism in complex statements. This was also shown in tweets that included unclear and implicit benevolent sexist notions, as they were misclassified as non-sexist.

Therefore, for the model to gain the ability to distinguish between the aforementioned contexts, we recommend increasing the size of the training data so that it includes more benevolent sexist and non-sexist statements in these specific contexts. In specific, the model needs to be trained on more long statements of both classes. In addition, non-sexist statements expressing marriage issues and sexist statements with implicit sexism notions should be added to further enhance the model’s classification performance.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this paper, we documented the construction of a representative corpus of sexism from quotes' websites, online articles, and the Google-Advanced Search tool. The corpus was annotated according to the Ambivalent Sexism Theory, which suggests that sexism has two subcomponents: hostile sexism and benevolent sexism. To minimize subjectivity and take into consideration the fact that benevolent sexism is a positively expressed type of sexism and is challenging to capture, the annotation was done by female and male students who have knowledge of gender issues.

The objective of this research was to create a data-centric system that allows for the detection of benevolent sexism in large amounts of data. Following the construction of the sexism corpus, we trained and tested nine classification machine learning models on these benevolent sexist and non-sexist statements. The models were evaluated using classification metrics including cross-validation accuracy, cross-validation AUC score, test AUC score, and the ROC Curve. Since the expression of benevolent sexism is seemingly positive and sometimes neutral, we were concerned about the model's ability to differentiate benevolent sexism from non-sexism, so we focused mainly on a high AUC score. The best performing model was the C-Support Vector Classification model. This model was further tuned to choose the best hyperparameter combination and then the classification threshold was tuned to meet our research objective of capturing the benevolent sexist statements and distinguishing between these statements and non-sexism.

As a means to validate our model's performance, we used the model to classify statements from a similar context (quotes) and a broader context (tweets on Mother's Day). In both validation experiments, the model's performance was consistent in detecting benevolent sexism. In specific, the model achieved significantly higher results in the classification of tweets. That is to say, the data on which the model was trained provided it with a high generalizability performance, which allowed us to meet our first objective of creating a representative benevolent sexism dataset.

However, we notice a model bias on special-context expressions such as marriage and the objective description of the physical characteristics of men and women. In addition, the model revealed a weakness in detecting implicit benevolent sexist expressions.

This reveals some of the potential limitations of our study. First, due to time constraints, we only focused on tuning one classification model. This made us skip exploring potential higher classification performance in other models. In addition, the annotators were able to label a dataset of 2,962 statements, which were used as training and testing sets. However, we faced difficulties in securing annotators to label our validation datasets, so the annotation of them was done by the author of the study, who has a fair knowledge of gender issues.

In future work, we plan on extending the training datasets to include more negative statements from the different contexts in which benevolent sexism is mostly present. We also aim at experimenting with deep neural networks to test for potential improvement in the model. In addition, we aim at executing the model into a system. This system can be utilized by users to check for benevolent sexism in text, and the positively classified sentences can be fed into the model to enhance its detection ability.

In a word, despite the seemingly benign expression of benevolent sexism, its impact on women and society is significant and needs more attention. The integration of concepts and knowledge from fields of social science, gender studies, and data science has become a must in a time where freedom of expression is being exploited.

REFERENCES

Abburi, H., Parikh, P., Chhaya, N., & Varma, V. (2020). Fine-grained multi-label sexism classification using semi-supervised learning. Web information systems engineering – WISE 2020 (pp. 531-547). Springer International Publishing. https://doi.org/10.1007/978-3-030-62008-0_37

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. Ithaca: Cornell University Library, arXiv.org. https://doi.org/10.1145/3041021.3054223

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances.

Barrientos, S., & Avendano, K. (2022). These International Women's Day Quotes Will Help You Unleash Your Inner Goddess. Good Housekeeping. https://www.goodhousekeeping.com/life/g26326977/international-womens-day-quotes/

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.

Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Bzdok, D., Krzywinski, M., & Altman, N. (2018). Machine learning: supervised methods. Nature Methods, 15(1), 5.

Casad, B. J., Salazar, M. M., & Macina, V. (2015). The real versus the ideal: Predicting relationship satisfaction and well-being from endorsement of marriage myths and benevolent sexism. Psychology of Women Quarterly, 39(1), 119-129. https://doi.org/10.1177/0361684314528304

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Paper presented at the 785-794. https://doi.org/10.1145/2939672.2939785

Chen, Z., Fiske, S. T., & Lee, T. L. (2009). Ambivalent sexism and power-related gender-role ideology in marriage. Sex Roles, 60(11-12), 765-778. https://doi.org/10.1007/s11199-009-9585-9

Claytor, T. (2022). How to Be a Lady (with Pictures) - wikiHow. wikiHow. https://www.wikihow.com/Be-a-Lady

Cross, E. J., Overall, N. C., & Hammond, M. D. (2016). Perceiving partners to endorse benevolent sexism attenuates highly anxious Women's negative reactions to conflict. Personality & Social Psychology Bulletin, 42(7), 923-940. https://doi.org/10.1177/0146167216647933

Dinh, A. (2019). Support Vector Machine (SVM). Retrieved August 27, 2022, from https://dinhanhthi.com/support-vector-machine

Dobilas, S. (2021). SVM classifier and RBF kernel‐how to make better models in Python. Medium. Retrieved August 27, 2022, from https://towardsdatascience.com/svm-classifier-and-rbf-kernel-how-to-make-better-models-in-python-73bb4914af5b

Elhamy, M. (2021). How to Be the Woman Every Man Wants to Marry - Cairo West Magazine. Cairo West Magazine. https://cairowestmag.com/how-to-be-the-woman-every-man-wants-to-marry/

Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter.

Journal of Intelligent & Fuzzy Systems, 36(5), 4743-4752. https://doi.org/10.3233/JIFS-179023

Friedman J, Hastie T, Tibshirani R, et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), 337–407.

Gilani, A. (2019). Machine Learning Basics: Support Vector Machines. Medium. Retrieved August 27, 2022, from https://medium.datadriveninvestor.com/machine-learning-basics-support-vector-machines-358235afb523

Goel, S., Madhok, R., & Garg, S. (2018). Proposing contextually relevant quotes for images. Advances in information retrieval (pp. 591-597). Springer International Publishing. https://doi.org/10.1007/978-3-319-76941-7_49

Gopalakrishnan, M. (2021). A brief history of Afghan women's rights | DW | 23.01.2022. DW.COM. https://www.dw.com/en/a-brief-history-of-afghan-womens-rights/a-60449450

Grosz, D., & Conde-Cespedes, P. (2020). Automatic detection of sexist statements commonly used at the workplace. Trends and applications in knowledge discovery and data mining (pp. 104-115). Springer International Publishing. https://doi.org/10.1007/978-3-030-60470-7_11

Guido, S., & Müller, A. C. (2016). Introduction to machine learning with python. O'Reilly Media.

Hammond, M. D., & Overall, N. C. (2013). When relationships do not live up to benevolent ideals: Women's benevolent sexism and sensitivity to relationship problems. European Journal of Social Psychology, 43(3), 212-223. https://doi.org/10.1002/ejsp.1939

Hammond, M. D., & Overall, N. C. (2015). Benevolent sexism and support of romantic Partner's goals: Undermining Women's competence while fulfilling Men's intimacy needs. Personality & Social Psychology Bulletin, 41(9), 1180-1194. https://doi.org/10.1177/0146167215593492

Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

Horbonos, P. (2020). Comparing a variety of Naive Bayes classification algorithms | by Pavlo Horbonos | Towards Data Science. Medium; towardsdatascience.com. https://towardsdatascience.com/comparing-a-variety-of-naive-bayes-classification-algorithms-fc5fa298379e

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of finnish text documents. In Proceedings of the Thirteenth ACM International Conference on Information And Knowledge Management (pp. 625-633).

Kotsiantis, S. B. (2011;2013;). Decision trees: A recent overview. The Artificial Intelligence Review, 39(4), 261-283. https://doi.org/10.1007/s10462-011-9272-4

Leaper, C., Gutierrez, B. C., & Farkas, T. (2022). Ambivalent sexism and reported relationship qualities in emerging adult heterosexual dating couples. Emerging Adulthood (Thousand Oaks, CA), 10(3), 776-787. https://doi.org/10.1177/2167696820934687

Lee, T. L., Fiske, S. T., Glick, P., & Chen, Z. (2010). Ambivalent sexism in close relationships: (hostile) power and (benevolent) romance shape relationship ideals. Sex Roles, 62(7-8), 583-601. https://doi.org/10.1007/s11199-010-9770-x

Liles, M. (2022). 100 International Women’s Day Quotes (2022) - Happy Women’s Day Quotes - Parade: Entertainment, Recipes, Health, Life, Holidays. Parade: Entertainment, Recipes, Health, Life, Holidays. https://parade.com/975103/marynliles/international-womens-day-quotes/

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1(4), 309-317.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. PloS One, 15(8), e0237861-e0237861. https://doi.org/10.1371/journal.pone.0237861

Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., & Varma, V. (2019). Multi-label categorization of accounts of sexism using a neural framework.

Park, J. H., & Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter.

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Detecting offensive language in tweets using deep learning. Ithaca: Cornell University Library, arXiv.org. https://doi.org/10.1007/s10489-018-1242-y

Porter, M. F. (2006). An algorithm for suffix stripping. Program: Electronic Library and Information Systems, 40(3), 211-218. https://doi.org/10.1108/00330330610681286

Quote Master (n.d.). Quote Master | Quotes about Everything. Retrieved September 3, 2022, from https://www.quotemaster.org/

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In Proceedings of the 20th International Conference on Machine Learning (ICML-03) (pp. 616-623).

Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence (Vol. 3, No. 22, pp. 41-46).

Samory, M., Sen, I., Kohne, J., Floeck, F., & Wagner, C. (2021). "call me sexist, but.": Revisiting sexism detection using psychological scales and adversarial samples. (). Ithaca: Cornell University Library, arXiv.org.

SPARCK JONES, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1), 11.

Sreekanth, V. (2018). 49 Best Wife Quotes To Touch Her Heart. WisdomTimes. https://www.wisdomtimes.com/blog/wife-quotes/

Sriram, K. (2021). Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2022 | upGrad blog. upGrad Blog; www.upgrad.com. https://www.upgrad.com/blog/multinomial-naive-bayes-explained

Support Vector Machines. (n.d.). Retrieved August 27, 2022, from Sklearn website: https://scikit-learn.org/stable/modules/svm.html

Veen, S. V., Cansfield, B., & Muir-Bouchard , S. (2018). Ten harmful beliefs that perpetuate violence against women and girls | Oxfam International. Oxfam International. https://www.oxfam.org/en/ten-harmful-beliefs-perpetuate-violence-against-women-and-girls

Zeinert, P., Inie, N., & Derczynski, L. (2021). Annotating online misogyny. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 3181-3197).