When does a Compliment become Sexist? Analysis and Classification of Ambivalent Sexism using Twitter Data

by

Akshita Jha, Radhika Mamidi

in

55th annual meeting of the Association for Computational Linguistics (ACL)
(*ACL-2017*)

Report No: IIIT/TR/2017/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
July 2017

When does a Compliment become Sexist? Analysis and Classification of Ambivalent Sexism using Twitter Data

Akshita Jha **Radhika Mamidi**
Kohli Center on Intelligent Systems (KCIS),
International Institute of Information Technology, Hyderabad (IIIT Hyderabad)
akshita.jha@research.iiit.ac.in, radhika.mamidi@iiit.ac.in

Abstract

Sexism is prevalent in today's society, both offline and online, and poses a credible threat to social equality with respect to gender. According to ambivalent sexism theory (Glick and Fiske, 1996), it comes in two forms: Hostile and Benevolent. While hostile sexism is characterized by an explicitly negative attitude, benevolent sexism is more subtle. Previous works on computationally detecting sexism present online are restricted to identifying the hostile form. Our objective is to investigate the less pronounced form of sexism demonstrated online. We achieved this by creating and analyzing a dataset of tweets that exhibit benevolent sexism. We classified tweets into 'Hostile', 'Benevolent' or 'Others' class depending on the kind of sexism they exhibit, by using Support Vector Machines (SVM), sequence-to-sequence models and FastText classifier. We achieved the best F1-score using FastText classifier. Our work aims to analyze and understand the much prevalent ambivalent sexism in social media.

1 Introduction

Sexism, as given by the Oxford dictionary, is the *'prejudice, stereotyping, or discrimination, typically against women, on the basis of sex'*. Sexism is rife in the society's belief system and its manifestation online is not uncommon (Eadicicco, 2014). For example, Australian game show, *My Kitchen Rules* often prompts sexist tweets against its female participants. E.g.: *'Trying to find something pretty about these blonde idiots. #MKR'*. However, evidence suggests that sexist remarks may not always express negative emotion (Becker and Wright, 2011). For instance, Rio Olympics shed light on the blatant as well as seemingly innocuous sexism that female athletes face, when, after the victory of 3-time Olympian Corey Cogdell-Unrein in women's trap shooting, Chicago Tribune tweeted, *'Wife of a Bears' lineman wins a bronze medal today in Rio Olympics'*[1]. Katie Ledecky's record breaking win in 400-meter freestyle race was applauded by a lot of people while simultaneously commenting that *'she swims like a man'*[2]. These are excellent examples of benign form of sexism prevailing in these times.

In their seminal paper, Glick and Fiske (1997) proposed ambivalent sexism theory that talked about two related but opposite orientations towards a particular gender: (i) Hostile Sexism (HS), *i.e.*, sexist antipathy and (ii) Benevolent Sexism (BS), *i.e.*, a subjectively positive view towards men or women. Hostile sexism is angry, harsh and expresses an explicitly negative viewpoint. E.g.: *'Jus gonna say it...again....DUMB BITCH! #MKR'*. Benevolent Sexism, on the other hand, is often disguised as a compliment. E.g.: *'They're probably surprised at how smart you are, for a girl'*. Moreover, there is a reverence for the stereotypical role of women as mothers, daughters and wives. BS puts women on a pedestal, but reinforces their sub-ordination. E.g.: *'No man succeeds without a good woman besides him. Wife or mother. If it is both, he is twice as blessed'*. Despite the positive feelings of BS, it's underpinnings lie in masculine dominance and stereotyping both men and women. It shares the common assumption that women inhabit restricted domestic roles and are the 'weaker sex'. Although, it may not be immediately apparent, this also implicitly stereotypes men.

Sexism has far-reaching consequences for women as well as men. It has been seen that despite it's seemingly positive and inoffensive tone,

[1] https://twitter.com/chicagotribune/status/762401317050605568
[2] https://tinyurl.com/y7zgsuyr

Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, pages 7–16,
Vancouver, Canada, August 3, 2017.

benevolent sexism has worse effects than hostile sexism on women's cognitive performance (Dardenne et al., 2007). Furthermore, the experiments conducted by Russo et al. (2014) demonstrate how social justification (Jost et al., 2004; Jost and Kay, 2005) and benevolent sexism are positively correlated. Additionally, they conclude that gender inequality is promoted not only by hostile sexism but also by the subtle and more deceptive, benevolent sexism.

Recently, efforts have been made for detection of sexist content from the internet. Some of the tweets in Waseem and Hovy's (2016) publicly available hate speech dataset of 16k tweets are sexist. But as expected in a hate speech corpus, these sexist tweets express only hostile sexism. It is evident that the approaches that detect sexism online have overlooked benevolent sexism.

In order to address the above shortcoming, we propose computational models to automatically classify a tweet into one of the three classes:

- *Benevolent*: if tweet exhibits subjectively positive sentiment but is sexist
- *Hostile*: if the tweet exhibits explicitly negative emotion and is sexist
- *Others*: if the tweet is not sexist

To the best of our knowledge, there has not been any previous study in computationally identifying benevolent sexism and classifying sexist content into two different classes depending on the nature of sexism.

The rest of the paper is organized as follows. Section 2 presents existing literature in related areas like hate speech detection, sentiment analysis and identification of sexist content from social psychology point of view. Section 3 illustrates the process of dataset creation and annotation for BS tweets. Additionally, it describes the available dataset of HS tweets that we used for our experiments. Section 4 and 5 describe the technical aspects of the experiments conducted for the classification of tweets. We discuss the results of the experiments in Section 6 before concluding the paper in Section 7.

2 Related Work

A considerable amount of work has been done in social psychology for identification of sexist content and its impact. Research has provided evidence that not only men but also women endorse sexist beliefs (Barreto and Ellemers, 2005; Glick et al., 2000; Jackman, 1994; Kilianski and Rudman, 1998; Swim et al., 2005). Becker and Wagner (2008) introduce Gender Identity Model (GIM) using social identity theory (SIT) (Hogg, 2016) and social role theory (SRT) (Eagly et al., 2000) to explain women's endorsement of sexist beliefs. They conclude that women reject benevolent and hostile sexism when they highly identify themselves with the category 'women' and have a progressive outlook. In contrast, gender role preference has weaker or no effect on sexist beliefs when women do not strongly identify themselves with their gender in-group.

The work by Bolukbasi et al. (2016) revealed the hidden gender bias in Word2Vec. They showed how Word2Vec word embeddings were sexist because of the bias in news articles that made up the Word2Vec corpus. For a relation like, *'father : doctor :: mother : x'*, Word2Vec gives x = *nurse*. And the query *'man : computer programmer :: woman : x'*, returns x = *homemaker*. In order to address this warping, they transformed the vector space using a method called 'hard de-biasing' and removed the bias.

Hate speech detection, that includes identification of sexist content, has garnered a lot of attention in recent times. Djuric et al. (2015) try to address this problem in online user comments. Using neural networks, they learn distributed low-dimensional text representations, where semantically similar comments and words reside in the similar part of vector space. They, then, feed this to a linear classifier to identify hateful and clean comments. Davidson et al. (2017) use hate speech lexicon to collect tweets containing hate speech keywords. They train a multi-class classifier to separate these tweets into one of the three classes: those containing hate speech, only offensive language, and those with neither. Hate speech dataset, containing sexist tweets, has been made publicly available by Waseem and Hovy (2016). This dataset contains 16k tweets that fall into one of the three classes: sexist, racist or neither. They list a set of criteria based on critical race theory to annotate the data and then use Support Vector Machines (SVM) with handcrafted features to classify tweets. However, one of the major drawbacks of the decsribed approaches and dataset is that it takes into account only hostile sexist tweets.

To better understand the nature of sexism, sentiment analysis can be done. In recent times, sentiment analysis of Twitter data has received a lot of attention (Pak and Paroubek, 2010). Some of the early works by Go et al. (2009) and Bermingham and Smeaton (2010) use distant learning to acquire sentiment data. They show that using unigrams, bigrams and part-of-speech (POS) tags as features, SVM outperforms other classifiers like Naive Bayes and MaxEnt. To remove the need for feature engineering, Agarwal et al. (2011) use POS-specific prior polarity features and tree kernel for sentiment analysis. To detect contextual polarity using phrase-level sentiment analysis, Wilson et al. (2005) identify whether a phrase is neutral or polar. If the phrase is polar, they then disambiguate the polarity of the polar expression. State-of-the-art sentiment analyzers use deep learning techniques like Convolutional Neural Network (CNN) (Dos Santos and Gatti, 2014) and Recursive Neural Network (Tang et al., 2015) based approach to learn features automatically from the input text.

3 Dataset

For the purpose of classification of tweets on the basis of the type of sexism, we required a dataset that displayed benevolent sexism (BS). Hence, we created our own corpus of tweets belonging to 'Benevolent' class. In addition to this, we used the publicly available hate speech corpus (Waseem and Hovy, 2016) to collect tweets belonging to 'Hostile' and 'Others' classes. Tweets labelled as 'sexist' and 'neither' in the hate-speech dataset make up the 'Hostile' and 'Others' class in our corpus respectively. Distribution of tweets in the combined corpus has been shown in Table 1.

	Total Tweets	**Unique Tweets**
Benevolent	7,205	712
Hostile	3,378	2,254
Others	11,559	7,129
Total	22,142	10,095

Table 1: Distribution of tweets in the combined corpus.

For creation of the Benevolent Sexist dataset, we collected a total of 95,292 tweets. Out of these, we manually identified 7,205 BS tweets (including retweets). This dataset is publicly available[3]. However, the total number of unique tweets identified, after removing retweets, were only 712 in number. The total number of tokens in the created dataset is 74,874. The mean length of BS tweets is 80.95, with a standard deviation of 25.75. The dataset also contains the metadata of each tweet, like username, time of creation of the tweet, it's geographic location, number of retweets and number of likes.

3.1 Data Collection

We collected data using the public Twitter Search API. The terms queried were common phrases and hashtags that are generally used when exhibiting benevolent sexism. Some of them were: *'as good as a man', 'like a man', 'for a girl', 'smart for a girl', 'love of a woman', '#adaywithoutwomen', '#womensday', '#everydaysexism' and '#weareequal'*. These lead to a dataset of tweets that were sexist in nature, both towards women and men. E.g.: *'He is a man who can't act like a man'* is sexist towards men. We extracted tweets that were in English. After we had manually identified benevolent tweets (explained is Section 3.2), we asked three 23-year old non-activist feminists to cross-validate the collected unique tweets to remove any kind of annotator bias. Fleiss' kappa score was calculated to assess the reliability of the agreement between the validators. It was found to be 0.74 which corresponds to 'substantial agreement' between the annotators (Fleiss et al., 1969).

3.2 Identification

To identify and annotate BS, we made use of the ambivalent sexism theory proposed by Glick and Fiske (1997) in social psychology. Sexism is hypothesized to encompass three sources of male ambivalence: *Paternalism, Gender Differentiation and Heterosexuality*. Each of these three components have two types, one of them results in hostile sexism and the other gives rise to benevolent sexism.

- *Paternalism*: Paternalism encompasses *dominative paternalism* and *protective paternalism*. Supporters of the former hold the view of women not being fully competent adults (Brehm, 1992; Peplau et al., 1983); whereas those who support the latter, view women as the weaker sex who need to be loved, cherished and protected (Peplau et al., 1983; Tavris et al., 1984). Protective paternalism

[3]Dataset: https://github.com/AkshitaJha/NLP_CSS_2017/

Paternalism	HS (Dominative)	: Women should stay at home.
	BS (Protective)	: Women are like flowers who need to be cherished!
Gender Differentiation	HS (Competitive)	: Women are incompetent at work.
	BS (Complementary)	: It's so good that I thought your brother wrote it!
Heterosexuality	HS (Hostility)	: I would like to fuck Kat, stupid slut!
	BS (Intimacy)	: What is man without the love of a woman!

Table 2: Examples tweets showing ambivalent sexism.

results in benevolent sexism whereas dominative paternalism results in hostile.

- *Gender Differentiation*: Akin to dominative paternalism, *competitive gender differentiation* justifies patriarchy in the society by viewing men as ones having governing capabilities in the society (Tajfel, 2010). This gives rise to hostile sexism. On the other hand, *complementary gender differentiation* results in benevolent sexism as it shows women having favourable traits that men stereotypically lack (Eagly and Mladinic, 1994).

- *Heterosexuality*: Similarly, *heterosexual intimacy* gives rise to benevolent sexism by viewing women as romantic objects with a genuine desire for psychological closeness (Berscheid et al., 1989); and *heterosexual hostility* is shown in cases where, for some men sexual attraction towards women may not be separate from the desire to dominate them (Bargh and Raymond, 1995; Pryor et al., 1995). This results in hostile sexism.

Table 2 shows some example tweets that highlight the ambivalent sexist attitude towards women. In order to clearly identify benevolent sexism, we studied the tweets and analyzed if it showed any one the three behaviors: *protective paternalism, complementary gender differentiation,* and *heterosexual intimacy*. If the tweet exhibited any one of the above, we annotated it as benevolently sexist.

3.3 Comparison of Hostile and Benevolent Sexist Tweets

The statistical difference in the distribution of hostile and benevolent sexist tweets in the combined dataset can be determined from Table 2. It is interesting to note that despite the total number of BS tweets (7,205) being almost double the total number of HS tweets (3,378), the number of unique BS tweets (712) is just one-third that of the unique HS tweets (2,254). Since benevolent sexism seems harmless, noble, and even romantic at times, it is retweeted more number of times as compared with tweets that exhibit hostile sexism.

Hostile	Benevolent
not	man
sexist	woman
#mkr	women
women	like
kat	#womensday
girls	love
like	good
call	girl
#notsexist	#adaywithoutwomen
female	without

Table 3: Most frequent content words in HS and BS tweets.

Table 3 shows the most common content words used in hostile and benevolent tweets. Apart from the words, 'girl(s)' and 'women', which are frequent in both kinds of tweets (as sexism is commonly expressed against females), we see that content words with high frequency differ significantly.

Hostile	Benevolent
kat and andre	think like man
sexist don't like	act like man
call sexist whatever	act like lady
sexist can't stand	last love man
blondes pretty faces	first love woman
dumb blondes pretty	love like woman
sexist hate female	without love woman
don't like female	lady think like
comedians aren't funny	man love like

Table 4: Most frequent tri-grams in HS and BS tweets.

Most frequent trigrams in hostile and benevo-

lent tweets are shown in Table 4. As hypothesized, benevolent tweets have trigrams that express positive attitudes while trigrams of hostile tweets express explicit negative attitude.

Table 5 illustrates the most frequent adjectives used for both hostile and benevolent tweets. We observe that frequent adjectives in HS tweets display a negative sentiment whereas adjectives in BS tweets display positive sentiment.

Hostile	Benevolent
dumb	real
hot	strong
bad	beautiful
stupid	better
awful	great

Table 5: Most frequent adjectives in HS and BS tweets.

All the above illustrations are in line with our hypothesis which states that sexism in the benevolent form is camouflaged as a compliment and is hence difficult to pinpoint; whereas, hostile sexism is evidently negative and can be easily identified as sexist.

3.4 Pre-processing

Pre-processing of tweets involved removal of usernames, punctuations, emoticons, hyperlinks/URLs and RT tag. Stop words were intentionally retained. The reason for this was that each tweet can contain a maximum of only 140 characters and removal of stop words would only lead to loss of information. For example in the tweet, *'Every guy should admit that #adaywithoutwomen is not a day worth living'*, stop word removal would remove *'not'* which as a result, would change a BS tweet to an HS tweet.

4 Methodology

For classification of tweets into one of the three classes: 'Benevolent', 'Hostile' and 'Others', we made use of machine learning techniques described below.

4.1 SVM

Support Vector Machines (SVM) (Cortes and Vapnik, 1995) are supervised learning models used for classification. To classify tweets in our dataset, we used term frequency-inverse document frequency (TF-IDF) (Salton and Buckley, 1988) as a feature, as it captures the importance of the given word in a document. TF-IDF is calculated as:

$$tfidf(t, d, D) = f(t, d) \times log\frac{N}{|\{d \epsilon D : t \epsilon d\}|}$$

where $f(t, d)$ indicates the number of times term, t appears in context, d and N is the total number of documents; $|\{d \epsilon D : t \epsilon d\}|$ represents the total number of documents where t occurs.

We ensure that SVM uses TF-IDF, to construct a separating hyperplane for given labelled training data and classify new tweets into one of the three classes: 'Benevolent', 'Hostile', or 'Others'. To find the optimal hyperplane, SVM tries to find a decision boundary that maximizes the margin by minimizing $||\mathbf{w}||$:

$$minf : \frac{1}{2}||\mathbf{w}||^2,$$

$$s.t. \quad y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1, \quad i = 1, ..., m$$

where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ is the input vector and b is the bias.

4.2 Sequence to Sequence model

A basic sequence-to-sequence model consists of an encoder and a decoder (Sutskever et al., 2014; Cho et al., 2014). For our experiment, we made use of a bi-directional RNN encoder-decoder (Schuster and Paliwal, 1997) with attention mechanism (Bahdanau et al., 2014) that employs Long Short Term Memory (LSTM) cells (Hochreiter and Schmidhuber, 1997) to modulate the flow of information. The encoder reads the input sequence and generates an intermediate hidden representation of fixed length, $\mathbf{c_o}$ given by:

$$\mathbf{c_o} = \sum_t \alpha_{ot}\mathbf{h_t}$$

where $\mathbf{h_t}$ denotes the hidden representation of $\mathbf{x_t}$, $\alpha_{ot} \epsilon [0, 1]$ and $\sum_t \alpha_{ot} = 1$. A learned alignment model computes the weight, $\alpha_o t$, for each $\mathbf{c_o}$ such that:

$$\alpha_{ot} = \frac{exp(e_o t)}{\sum_{t'} exp(e_o t')}$$

$$e_o t = a(\mathbf{s_{o-1}}\mathbf{h_t})$$

where $\mathbf{s_o}$ is the output of a recurrent hidden layer and $a(.)$ is a feed-forward neural network that

computes $\mathbf{h_t}$. The decoder then maps the intermediate representation into either one of the 'Benevolent', 'Hostile' or 'Others' class by computing:

$$P(y_1, .., y_O|\mathbf{x_1}, .., \mathbf{x_T}) = \prod_{o=1}^{O} P(y_o|y_1, .., y_{o-1}, \mathbf{c_o})$$

where lengths of the output and the input are O and T respectively. The posterior probability, y_o is calculated as:

$$P(y_o|y_1, .., y_{o-1}, \mathbf{c_o}) = g(\mathbf{y_o}, \mathbf{s_o}, \mathbf{c_o})$$

where $\mathbf{y_o}$ is the vector representation of y_o, *i.e.,* a one-hot vector followed by neural projection layer for dimension reduction and $g(.)$ is a softmax function.

4.3 FastText

FastText classifier, made available by Facebook AI Research has proven to be efficient for text classification (Joulin et al., 2016). It is often at par with deep learning classifiers in terms of accuracy, and much faster for training and evaluation. FastText uses bag of words and bag of n-grams as features for text classification. Bag of n-grams feature captures partial information about the local word order. FastText allows update of word vectors through back-propagation during training allowing the model to fine-tune word representations according to the task at hand (Bojanowski et al., 2016). The model is trained using stochastic gradient descent and a linearly decaying learning rate.

5 Experiments and Results

Experiments conducted for classification of tweets have been described below. We trained and tested our algorithm only on unique tweets to avoid learning any kind of bias from retweets. For evaluating the experiments, we use precision, recall and f-measure.

5.1 Polarity Detection

To detect the polarity of each tweet, we experimented with rule-based sentiment analysis techniques using linguistic features. First, using the Penn Treebank tagset (Marcus et al., 1993), all tweets were tagged for part-of-speech (POS). After this, we used the Stanford Shallow Parser (Pradhan et al., 2004) to chunk tweets and get all the phrases. We calculated the positive score and the negative score for each phrase in the tweet, using SentiWordNet (Baccianella et al., 2010) and subjectivity lexicon (Taboada et al., 2011). The overall sentiment score of a tweet was calculated by summing up the individual score of the phrases in the tweet. If this overall sentiment score of the tweet was greater than 0, then the tweet was marked as positive; if the overall sentiment score was less than 0, it was marked as negative; else the tweet was marked as neutral. Table 6 shows the results of the basic sentiment analysis of tweets.

	Hostile	**Benevolent**	**Others**
Positive	3.07%	83.06%	7.34%
Negative	86.48%	2.77%	15.72%
Neutral	10.45%	14.17%	76.94%

Table 6: Sentiment Analysis of tweets in the dataset.

5.2 SVM

For the purpose of our experiment, we used TF-IDF as a feature for SVM to classify previously unseen tweets into one the three classes. We implemented SVM using scikit (Pedregosa et al., 2011) library. Table 7 shows the precision, recall and F1-score after performing 10-fold cross validation.

5.3 Sequence to Sequence model

The implementation of the described Sequence to Sequence model has been done using tf-seq2seq framework (Britz et al., 2017) for Tensorflow (Abadi et al., 2016). The experiment was conducted after splitting the training set and the test set in the ratio 7 : 3. For 1000 epochs, with a batch-size of 32, the precision, recall and F1-score have been shown in Table 7.

5.4 FastText

The training set and the test set were split in 7 : 3 ratio for FastText. Table 8 reports precision at 1 of running FastText, using 100 dimension word vectors, for 5, 8, 10 and 15 epochs with a learning rate of 0.1 and the size of context window as 5. It is observed that there is no improvement in the F1-score after 10 epochs.

	SVM			Seq2Seq		
	P	**R**	**F1**	**P**	**R**	**F1**
Benevolent	0.97	0.69	**0.80**	0.69	0.77	0.73
Hostile	0.89	0.33	0.48	0.57	0.65	**0.61**
Others	0.80	0.99	**0.89**	0.91	0.87	0.88

Table 7: Comparision of Precision (P), Recall (R) and F1 score (F1) of classification of tweets into HS, BS and Others class using SVM and Seq2seq models.

Epochs	Precision	Recall	F1-Score
5	0.81	0.81	0.81
8	0.84	0.84	0.84
10	0.87	0.87	0.87
15	0.87	0.87	0.87

Table 8: FastText Prec@1 for different epochs.

6 Discussion

Using basic linguistic features, rule-based polarity detection of tweets show that benevolent sexism have positive polarity whereas the tweets exhibiting hostile sexism have a negative polarity. This is in accordance with our hypothesis which states that benevolent sexism expresses a positive outlook, in contrast to hostile sexism that displays negative emotion.

For the purpose of classification of tweets into 'Benevolent', 'Hostile' or 'Others' class, Support Vector Machines (SVM) and Sequence to Sequence (Seq2Seq) classifier were implemented for baseline experiments. In SVM, the precision for the 'Benevolent' and 'Hostile' class is unusually high whereas the recall, specifically for the 'Hostile' class, is quite low. This implies that only 69% of BS tweets and 33% of HS tweets of the previously unseen test set have been labelled correctly. On comparing this with the results of Seq2Seq model, we observe that although the precision for classification of tweets into 'Benevolent' and 'Hostile' is not as high as that of SVM, the recall is 77% and 65% respectively for the two classes, which is better than the recall achieved using SVM. Seq2Seq takes into account the structure of the tweet, unlike the TF-IDF feature used in SVM, which is invariant to word order. This results in better recall.

The number of tweets in 'Others' class is significantly more than the number of tweets in 'Hostile' and 'Benevolent' classes combined. The performance of SVM and Sequence to Sequence models is known to improve, as the size of varied training data increases. This is further reflected in the high precision, recall and the comparable F1-score achieved for the 'Others' class using the two models.

Overall, SVM gives a slightly better F1-score for 'Benevolent' and 'Others' class, whereas Sequence to Sequence classifier performs better for 'Hostile' class. FastText outperforms both the above classifiers, with an F1-score of 0.87 for Prec@1. Since, a tweet has limited number of characters and may not exhibit long range dependencies, the word order of a tweet is successfully captured by FastText, by using it's bag of n-gram feature. This, combined with the fact that FastText has lesser number of parameters to tune, results in it's better performance than the proposed Seq2Seq model.

7 Conclusion and Future Work

We presented a detailed analysis for detection and classification of sexism in twitter data by building a combined corpus of benevolent and hostile sexist tweets. Using ambivalent sexism theory, we annotated tweets that showed sexism in the benevolent form. A limitation of our approach was that the method of gathering benevolently sexist tweets was biased towards the initial search terms and likely missed many forms of benevolent sexism. In future, we aim to address this concern by increasing the size of the dataset, using the aforementioned ambivalent sexism theory, while additionally solving the issue of the comparatively lesser number of unique benevolently sexist tweets. We also plan to take into consideration the gender of the user, the geographic location of a tweet and its length as features for future experiments.

Apart from understanding and identifying various kinds of sexism, the created dataset can additionally be used to recognize and analyze the events that trigger sexism online. The methods described can also be used in contexts outside of

social media, such as within workplace communications as means for automated assessment and eventual intervention. While the problem is far from solved, our experiments can be treated as a baseline for future work.

Our work is a step towards building a gender-neutral society. The insights derived from the analysis and experiments presented in this paper may prove beneficial in understanding the prevalence of ambivalent sexism in social-media data and serve as a starting point for more work in this field.

Acknowledgments

We thank the annotators and the three anonymous reviewers for their useful comments.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* .

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics, pages 30–38.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*. volume 10, pages 2200–2204.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

John A Bargh and Paula Raymond. 1995. The naive misuse of power: Nonconscious sources of sexual harassment. *Journal of Social Issues* 51(1):85–96.

Manuela Barreto and Naomi Ellemers. 2005. The perils of political correctness: Men's and women's responses to old-fashioned and modern sexist views. *Social Psychology Quarterly* 68(1):75–88.

Julia C Becker and Ulrich Wagner. 2008. Doing gender differently–. *Womens Internalization of Sexism: Predictors and Antidotes* page 51.

Julia C Becker and Stephen C Wright. 2011. Yet another dark side of chivalry: Benevolent sexism undermines and hostile sexism motivates collective action for social change. *Journal of personality and social psychology* 101(1):62.

Adam Bermingham and Alan F Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, pages 1833–1836.

Ellen Berscheid, Mark Snyder, and Allen M Omoto. 1989. The relationship closeness inventory: Assessing the closeness of interpersonal relationships. *Journal of personality and Social Psychology* 57(5):792.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*. pages 4349–4357.

Sharon S Brehm. 1992. *Intimate relationships*. Mcgraw-Hill Book Company.

D. Britz, A. Goldie, T. Luong, and Q. Le. 2017. Massive Exploration of Neural Machine Translation Architectures. *ArXiv e-prints* .

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

Benoit Dardenne, Muriel Dumont, and Thierry Bollier. 2007. Insidious dangers of benevolent sexism: consequences for women's performance. *Journal of personality and social psychology* 93(5):764.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009* .

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, pages 29–30.

Cícero Nogueira Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*. pages 69–78.

Lisa Eadicicco. 2014. This female game developer was harassed so severely on twitter she had to leave her home. *http://www.businessinsider.com/brianna-wu-harassed-twitter-2014-10?IR=T, Oct.* .

Alice H Eagly and Antonio Mladinic. 1994. Are people prejudiced against women? some answers from research on attitudes, gender stereotypes, and judgments of competence. *European review of social psychology* 5(1):1–35.

Alice H Eagly, Wendy Wood, and Amanda B Diekman. 2000. Social role theory of sex differences and similarities: A current appraisal. *The developmental social psychology of gender* pages 123–174.

Joseph L Fleiss, Jacob Cohen, and BS Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 72(5):323.

Peter Glick and Susan T Fiske. 1996. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology* 70(3):491.

Peter Glick and Susan T Fiske. 1997. Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of women quarterly* 21(1):119–135.

Peter Glick, Susan T Fiske, Antonio Mladinic, José L Saiz, Dominic Abrams, Barbara Masser, Bolanle Adetoun, Johnstone E Osagie, Adebowale Akande, Amos Alao, et al. 2000. Beyond prejudice as simple antipathy: hostile and benevolent sexism across cultures. *Journal of personality and social psychology* 79(5):763.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1(12).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Michael A Hogg. 2016. Social identity theory. In *Understanding Peace and Conflict Through Social Identity Theory*, Springer, pages 3–17.

Mary R Jackman. 1994. *The velvet glove: Paternalism and conflict in gender, class, and race relations*. Univ of California Press.

John T Jost, Mahzarin R Banaji, and Brian A Nosek. 2004. A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political psychology* 25(6):881–919.

John T Jost and Aaron C Kay. 2005. Exposure to benevolent sexism and complementary gender stereotypes: consequences for specific and diffuse forms of system justification. *Journal of personality and social psychology* 88(3):498.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* .

Stephen E Kilianski and Laurie A Rudman. 1998. Wanting it both ways: Do women approve of benevolent sexism? *Sex roles* 39(5):333–352.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics* 19(2):313–330.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*. volume 10.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(Oct):2825–2830.

Letitia Anne Peplau et al. 1983. Roles and gender. *Close relationships* pages 220–264.

Sameer S Pradhan, Wayne H Ward, Kadri Hacioglu, James H Martin, and Daniel Jurafsky. 2004. Shallow semantic parsing using support vector machines. In *HLT-NAACL*. pages 233–240.

John B Pryor, Janet L Giedd, and Karen B Williams. 1995. A social psychological model for predicting sexual harassment. *Journal of Social Issues* 51(1):69–84.

Silvia Russo, Filippo Rutto, and Cristina Mosso. 2014. Benevolent sexism toward men: Its social legitimation and preference for male candidates. *Group Processes & Intergroup Relations* 17(4):465–473.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24(5):513–523.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Janet K Swim, Robyn Mallett, Yvonne Russo-Devosa, and Charles Stangor. 2005. Judgments of sexism: A comparison of the subtlety of sexism measures and sources of variability in judgments of sexism. *Psychology of Women Quarterly* 29(4):406–411.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37(2):267–307.

Henri Tajfel. 2010. *Social identity and intergroup relations*. Cambridge University Press.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*. pages 1422–1432.

Carol Tavris, Carole Wade, and Carole Offir. 1984. *The longest war: Sex differences in perspective*. Harcourt.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL-HLT*. pages 88–93.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, pages 347–354.