Laurence Breyer
Benjamin Byle
Emma Dahl
Francisco Fortes
Eleanor Manley
Tamara Qassem

Artificial General Intelligence Group Assignment

I. Business Problem

Sexual harassment is an immense problem within companies: over 60% [1]of women and 35%[2] of men are sexually harassed each year within the workplace but 99.8%[3] of them never report it. Sexual harassment is defined as 'any act perceived as discriminatory based on sex that makes an individual uncomfortable'[4] and doesn't necessarily have to be overtly sexual in nature, rather reflecting a degrading attitude based on sex and gender. We can characterise sexual harassment into three main categories[5]:

1. Gender harassment - an objectifying behaviour meant to demean an individual based on their sex.
2. Unwarranted sexual attention - an unwelcome, sexually intentioned behaviour directed towards a person.
3. Sexual coercion - pressure or force upon another to participate in sexual activities.

Sexual harassment within the workplace is a very troubling occurrence because it implies the behaviour may have to be endured for an individual to maintain their job, their position within the company, and creates a hostile environment in which a person will not be able to perform their professional activities to the best of their abilities. The loss of productivity within a company from employees dealing with sexual harassment on an individual basis is huge[6]: 36% of employees who have experienced sexual harassment at work have trouble concentrating, 32% do not complete their tasks, and 17% make excuses to not show up at all to the workplace.

However, the effects of sexual harassment aren't merely significant on an individual level - companies' bottom lines are directly paying the costs for undealt sexual harassment cases[7]: $2.6 billion dollars are lost in productivity every year due to sexual harassment cases at work.

---

[1] Institut D'études Opinion Et Marketing En France Et À L'international, *Observatoire Européen Du Sexisme Et Du Harcèlement Sexuel Au Travail,* Ifop, (October 2019): p. 1
[2] Metta Space, *What are the effects of Sexual Harassment within the Worplace?*, (March 2020).
[3] Carly McCann, Donald Tomaskovic-Devey, & M.V. Lee Badgett, *Employer's Responses to Sexual Harassment,* Center for Employment Equity, University of Massachusetts, (Dec. 2018).
[4] Metta Space, *Definition of Sexual Harassment*, (May 2020).
[5] Based on: U.S. Merit Systems Protection Board Office of Policy and Evaluation, *Research Brief: Update on Sexual Harassment in the Federal Workplace*, (March 2018).
[6] Metta Space, *What are the effects of Sexual Harassment within the Worplace?*, (March 2020).
[7] Based on: Deloitte, *The economic costs of sexual harassment in the workplace,* (March 2019).

This is caused by a 32% increase in staff turnover, a 28% increase in absenteeism and a 24% loss in managerial time dealing with sexual harassment cases on an individual basis. Furthermore, if a sexual harassment case is leaked, the reputational costs are huge. Last year, $4 billion dollars were lost in reputational costs to companies that had not dealt efficiently with sexual harassment.

Sexual harassment is a colossal problem within institutions, yet no one is dealing with it. As such, in this paper we will propose an artificial intelligence (AI) system to alleviate the problem of sexual harassment within companies. We will bring to light how a natural language processing system (NLP) that would be installed in an enterprise would be capable of flagging up sexual harassment to go through companies' in-house correspondances.

II. Cognitive and Affective Functions

Before we dive into NLP and how sexual harassment can be detected by an algorithm, we first want to introduce the cognitive and affective functions of language. Language allows storing of complex information and is a double-edged sword in terms of processing. In order to build an artificial intelligence program around language processing, first we must discuss how language works. Studies such as Hurlburt's 1990 experiment[8] indicate that language is not a result of thinking, but in fact the active process of thinking. That's to say, we use language to develop thoughts, rather than simply having thoughts and then use language to express those thoughts. However, many members of the cognitive science community believe[9] that language is just an input-output system - a mechanism for communicating thoughts. This school of thought suggests that language is only one part of a larger picture in terms of understanding. For example, images, rather than language, is the best mechanism to understand shapes and colors. While highly visual things can be described in words, an image may be more effective to communicate meaning.

It was long believed that reading was a simple graph to sound decoding mechanism activating only a limited amount of the brain's areas. However, studies reveal that the neural mechanism behind the skill called "reading" is far more complicated than that: brain functions change during reading and activate sub-processes that enable reading to become comprehension.[10] All of those sub-processes, including sensory visual processing of the symbols, working and long-term memory, motor processing and understanding all take place in different areas of the brain. The more complicated the reading task becomes, the more regions will be activated. Reading one word takes place in Broca's area together with the frontal gyrus and the insular cortex. The reading of an entire sentence will encompass the activation of more areas[11]. In short, it is the relationship between word identification and the sub-processes associated with reading that matter. Reading, cognition, emotion, learning, and memory work together to promote the comprehensive understanding of the symbols that pass before your eyes.

---

[8] Russell Hurlburt, Plenum Press, *Sampling normal and schizophrenic inner experience* (1990).
[9] Behavioral and Brain Sciences, *The Cognitive Functions of Language* (December 2003).
[10] Buchweitz, A., Robert A. M., Leda M.B., & Marcel, A.J. (2009). *Brain Activation for Reading and Listening Comprehension: an fMRI Study of Modality Effects and Individual Differences in Language Comprehension*, Psychology Neuroscience
[11] Perfetti, C.A. & Gwen A. F., (2008). The Neural Bases of Text and Discourse Processing. Handbook of the Neuroscience of Language. Amsterdam: Academic press.

Furthermore, not just the individual areas matter but also the neural pathways that connect them.

For the sake of this paper, we will focus on the areas that are cornerstones to the understanding of written language. The regions with the largest roles are Broca's area (the posterior part of the left inferior frontal gyrus) and the Wernicke's area (posterior part of the superior temporal gyrus). In addition, six more areas are actively involved at the same time: the anterior cingulate gyrus, the prefrontal cortex, the basal temporal language area or fusiform gyrus, the cerebellum, the right hemisphere, and the elements of the limbic system.[12] The above-mentioned areas are all located in the left hemisphere of the brain. That is just for the reading of letters and words.

Nonetheless, words have different meanings in different contexts and discourse comprehension is key to understanding language semantics and when narratives, metaphors, inferences, references and so forth come into play. As such, the brain also needs the right hemisphere to give correct interpretations. The right hemisphere supports a broader range of meanings and associations[13]. Metaphor production, for example, is associated with the left angular gyrus and the posterior cingulate cortex. It is however, the brain activation between the angular gyrus and the right middle temporal gyrus that increases the creative qualities of metaphors understanding and responses[14].

For proper discourse comprehension, episodic future thought is involved as well. This suggests that there must be a flexible adaptation of the semantic memory (anterior temporal lobe) to further enhance this process. Memory, concept generation/integration, abstract and thematic relationships, storyline structure, story construction and emotional valuation are behind the set up of this advanced comprehension.[15] The limbic system is partially responsible for that which is, together with other paralimbic regions, closely associated with the hypothalamus and brainstem nuclei.

Therefore, the processing of language in our brain is not just a connection of individual series of events but an extremely complicated interconnected cooperation of many different brain areas and neural pathways. Despite language favouring the left hemisphere of our brain, for exhaustive interpretation, we need the right hemisphere as well. These processes happen in the left and right hemispheres individually, but it is clear that there is an exchange of information and processes between them as well.

The significance of language and its role in thinking is widely debated in the cognitive field. One such thesis suggests that the brain, in addition to having processing centers for vision,

---

[12] Lem, L. *Beyond Broca's and Wernicke's Areas: a new perspective on the neurology of language, Issues in Applied Linguistics*, 2, 2(1992).

[13] Frishkoff G.A., *Hemispheric differences in strong versus weak semantic priming: Evidence from event-related brain potentials*. Brain & Language, 100 (1), (2007): p. 23–43

[14] Fink et al., (2009) *Creative brain: investigation of brain activity during creative problem solving by means of EEG and fMRI Hum. Brain Mapp*., 30 (2009)

[15] Hruby, G.G. *Grounding reading comprehension in the neuroscience literature*. In Susan E.I. & Gerald G. D. (Eds.) Handbook of the Research of Reading Comprehension, New York: Routledge (2009) : p.189-224

hearing, face recognition, etc., also has conceptual processing centers. A study[16] conducted with lab rats and humans indicates the intersection of language and understanding. Addressed by Carruthers but better summarized by RadioLab's Hidden Brain podcast[17], the experiment showed rats an item in a room, and then disoriented them and asked them to find the item. The room had various features such as walls with different colors, patterns, scents, and dimensions. While the rats were able to distinguish colors, scents, etc., they weren't able to process spatial or relative information- for example, if the item is on the left or right side of the pink wall. A similar test recreated with humans asked them to find items in a room where the location included a complex thought- ie, "the item is on the left side of the blue wall and behind the green lamp". All the subjects were able to locate the item, but when the experiment was re-run and subjects had to simultaneously repeat back a passage being read to them, they were unable to find the items. These two experiments suggest that language allows us to develop complex meaning, and understand conceptual ideas that involve several processing centers simultaneously - Carruthers refers to this as "cross- modular thinking"[18]. While language isn't the only component necessary to understand and develop complex thoughts, it is the string that ties together different cognitive centers.

In order to solve the problem of sexual harassment in the workplace, we are proposing a natural language processing algorithm that will identify whether a message has a sexual connotation or not. While humans learn language by mimicking what is spoken around them, machines learn mostly through grammar rules and by examples fed through the machine.

III. Theoretical Blueprint & Proof of Concept

We decided to build a model that would be able to detect the words or phrases used within workplace correspondances and determine whether they are considered as abusive or offensive within the sexual harassment sphere.

In order to build a model that detects abusive language and sexual harassment our team has decided to use a Dataset from *Automated Hate Speech Detection and the Problem of Offensive Language* by Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber[19]. The dataset consists of tweets and we are aware that they are not workplace correspondences but we believe correspond best to colloquial terms that individuals may apply when speaking to each other. Thedataset has 24,783 rows and it is composed by 7 columns :

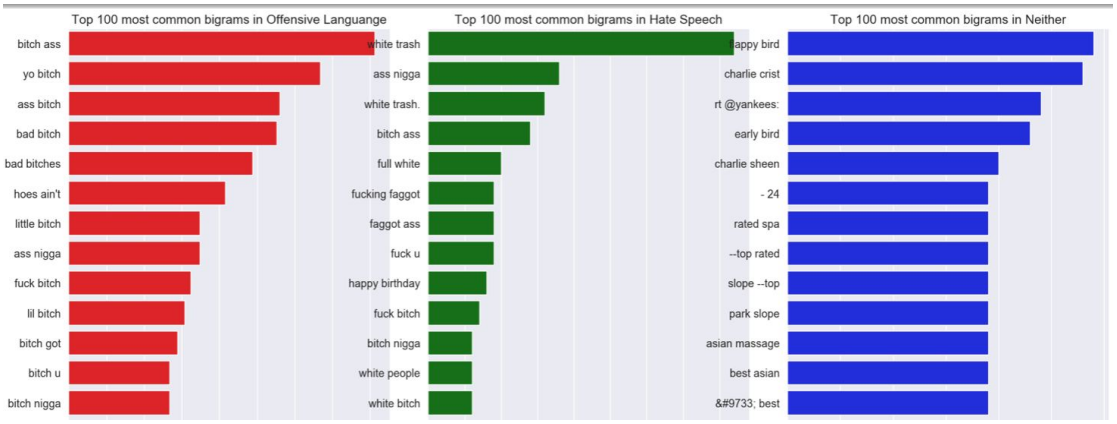
- count = number of CrowdFlower users who coded each tweet (min is 3, sometimes more users coded a tweet when judgments were determined to be unreliable by CF).
- hate_speech = number of CF users who judged the tweet to be hate speech.
- offensive_language = number of CF users who judged the tweet to be offensive.

[16] Cheng, K, Cognition 2. *A purely geometric module in the rat's spatial representation* (1986).
[17] Hidden Brain. *Lost In Translation: The Power Of Language To Shape How We View The World.* (2018)
[18] Peter Carruthers, Behavioral and Brain Sciences, *The Cognitive Functions of Language* (December 2003).
[19] Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber, *Automated Hate Speech Detection and the Problem of Offensive Language,* Cornell University, (2017).

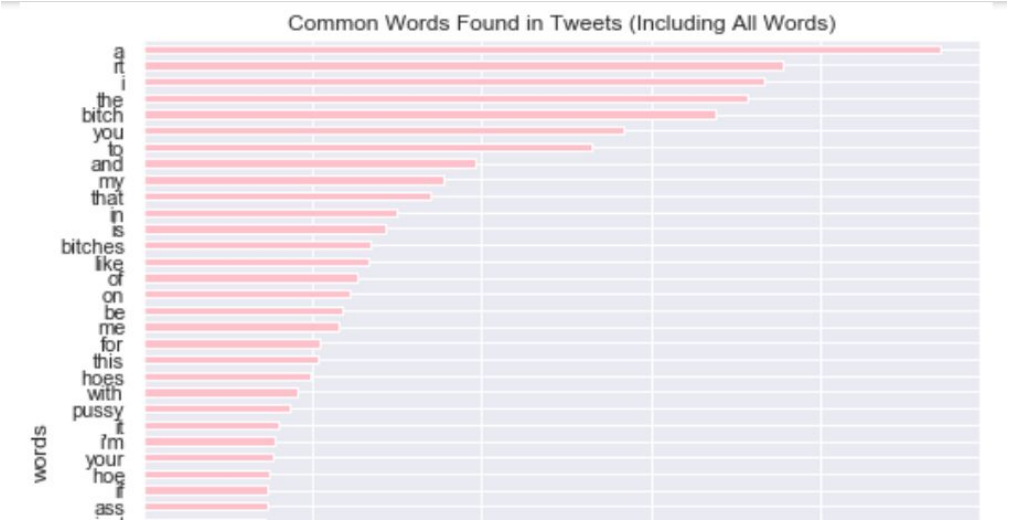
- neither = number of CF users who judged the tweet to be neither offensive nor non-offensive.
- class = class label for majority of CF users. 0 - hate speech 1 - offensive language 2 - neither
- Tweet = actual tweet text
- Unnamed:0 = row number

Using the above mentioned dataset, some basic exploratory data analysis was performed on tweets. Luckily, this dataset contained 0 nulls which made our job much easier when it came to data cleaning. All in all, we checked the number of words, stop words, characters, punctuation, hashtags, mentions and finally URLs. Considering the texts analyzed are real tweets implied the presence of its own 'lingo' which we had to take into consideration. Therefore, this gave us some of our features to input in our model.

Next, we performed an N-grams analysis, including unigrams, bigrams and trigrams, in order to get a clearer idea of what our offensive, hate speech and neither categories look like. Below is an image of only some of the most common bigrams portraying astonishing results.



Additionally, we decided to check what the most common words are. The image below is just a snippet of some of the words that appeared. As you can see, most of these words are in fact stop words, with a combination of profanity, which is what inspired us to do some text cleaning.



In regards to the text preprocessing stage, we received the best results by removing the URLs, any extra symbols and numbers, punctuations, whitespaces, stop words, and finally emojis. Eliminating these displayed the tweets in a much clearer manner allowing us to finally begin modelling.

Now that we have brought to light a few insights from this data analysis and, taking into account the data cleaning that we just mentioned, we decided to do an ensemble of various models to see the one the best fits our needs. Using the n-grams we created previously we decided to use in a range of 1 to 2 grams, as it was the one that gave us our highest score. In our ensemble of models we started off our model building with the Naive Bayes classifier, and them we used Stochastic Gradient Descent, Logistic Regression, Decision trees, Random Forest with and without max tuning, K-Nearest Neighbours and Support Vector Machine. Ultimately we got the scores as you can see from the image below

	test_score	train_score	clf
SVC_poly3	0.890861	0.936094	SVC(C=1.0, break_ties=False, cache_size=200, c...
TF-IDF	0.88602	0.946888	SGDClassifier(alpha=0.0001, average=False, cla...
Naive_Bayes	0.880775	0.968274	MultinomialNB(alpha=1.0, class_prior=None, fit...
SVC_rbf	0.872705	0.963634	SVC(C=1.0, break_ties=False, cache_size=200, c...
LOG	0.872302	0.918894	LogisticRegression(C=1, class_weight=None, dua...
LogisticRegression	0.863627	0.993645	LogisticRegression(C=100, class_weight=None, d...
DecisionTreeClassifier	0.862417	0.998689	DecisionTreeClassifier(ccp_alpha=0.0, class_we...
KNN_n1	0.782933	0.998285	KNeighborsClassifier(algorithm='auto', leaf_si...
KNN_n3	0.779302	0.805659	KNeighborsClassifier(algorithm='auto', leaf_si...
KNN_n5	0.777285	0.788056	KNeighborsClassifier(algorithm='auto', leaf_si...
RandomForestClassifier_maxDepth25	0.775066	0.780894	(DecisionTreeClassifier(ccp_alpha=0.0, class_w...
RandomForestClassifier_maxDepth20	0.77325	0.77696	(DecisionTreeClassifier(ccp_alpha=0.0, class_w...
RandomForestClassifier_maxDepth15	0.773048	0.775799	(DecisionTreeClassifier(ccp_alpha=0.0, class_w...
RandomForestClassifier_maxDepth10	0.771434	0.775093	(DecisionTreeClassifier(ccp_alpha=0.0, class_w...
RandomForestClassifier_maxDepth20_maxFeatures5	0.771233	0.775093	(DecisionTreeClassifier(ccp_alpha=0.0, class_w...
RandomForestClassifier_maxDepth20_maxFeatures11	0.771233	0.775093	(DecisionTreeClassifier(ccp_alpha=0.0, class_w...
RandomForestClassifier_maxDepth20_maxFeatures9	0.771233	0.775093	(DecisionTreeClassifier(ccp_alpha=0.0, class_w...
RandomForestClassifier_maxDepth20_maxFeatures7	0.771233	0.775093	(DecisionTreeClassifier(ccp_alpha=0.0, class_w...
RandomForestClassifier_maxDepth20_maxFeatures1	0.771233	0.775093	(DecisionTreeClassifier(ccp_alpha=0.0, class_w...
RandomForestClassifier_maxDepth20_maxFeatures3	0.771233	0.775093	(DecisionTreeClassifier(ccp_alpha=0.0, class_w...

As you may see our highest scoring model was the Support Vector Machines with a score on our test set of 89%.On this model we adjusted the SVM classifier features by adding 'kernel = poly' and 'degree = 1'.

Ideally to further develop our model we would gather our own tweets specifically targeted towards sexual harassment therefore this would provide us with an even more accurate model to contributing to a solution for our business model. We would also like to see this model to be used in companies in order to detect sexual harassment and avoid this from happening within businesses.

IV. Conclusion

All in all, we have brought to light the huge problem that is sexual harassment within the workplace. Our ultimate goal is to bring to market a NLP that can understand workplace correspondances and detect when inappropriate messages are being sent. We have illustrated the affective and cognitive functions linked to reading and writing that need to be understood before creating a NLP model. From that point forward, our paper presents the blueprint of how we would tackle this issue as well as the results that we achieved within our Proof of Concept.

Bibliography:

- Behavioral and Brain Sciences, *The Cognitive Functions of Language* (December 2003).
- Buchweitz, A., Robert A. M., Leda M.B., & Marcel, A.J. (2009). *Brain Activation for Reading and Listening Comprehension: an fMRI Study of Modality Effects and Individual Differences in Language Comprehension*, Psychology Neuroscience
- Carly McCann, Donald Tomaskovic-Devey, & M.V. Lee Badgett, *Employer's Responses to Sexual Harassment,* Center for Employment Equity, University of Massachusetts, (Dec. 2018).
- Cheng, K, Cognition 2. *A purely geometric module in the rat's spatial representation* (1986).
- Deloitte, *The economic costs of sexual harassment in the workplace,* (March 2019).
- Fink et al., (2009) *Creative brain: investigation of brain activity during creative problem solving by means of EEG and fMRI Hum. Brain Mapp*., 30 (2009)
- Frishkoff G.A., *Hemispheric differences in strong versus weak semantic priming: Evidence from event-related brain potentials*. Brain & Language, 100 (1), (2007): p. 23–43
- Hidden Brain. *Lost In Translation: The Power Of Language To Shape How We View The World.* (2018)
- Hruby, G.G. *Grounding reading comprehension in the neuroscience literature*. In Susan E.I. & Gerald G. D. (Eds.) Handbook of the Research of Reading Comprehension, New York: Routledge (2009) : p.189-224
- Institut D'études Opinion Et Marketing En France Et À L'international, *Observatoire Européen Du Sexisme Et Du Harcèlement Sexuel Au Travail,* Ifop, (October 2019): p. 1
- Lem, L. *Beyond Broca's and Wernicke's Areas: a new perspective on the neurology of language, Issues in Applied Linguistics*, 2, 2(1992).
- Metta Space, *Definition of Sexual Harassment*, (May 2020).
- Metta Space, *What are the effects of Sexual Harassment within the Worplace?*, (March 2020).
- Perfetti, C.A. & Gwen A. F., (2008). The Neural Bases of Text and Discourse Processing. Handbook of the Neuroscience of Language. Amsterdam: Academic press.
- Peter Carruthers, Behavioral and Brain Sciences, *The Cognitive Functions of Language* (December 2003).
- Peter Carruthers, Behavioral and Brain Sciences, *The Cognitive Functions of Language* (December 2003).
- Russell Hurlburt, Plenum Press, *Sampling normal and schizophrenic inner experience* (1990).
- Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber, *Automated Hate Speech Detection and the Problem of Offensive Language,* Cornell University, (2017).
- U.S. Merit Systems Protection Board Office of Policy and Evaluation, *Research Brief: Update on Sexual Harassment in the Federal Workplace*, (March 2018).