GRADUATE CERTIFICATE IN INTELLIGENT REASONING SYSTEMS PRACTICE MODULE REPORT

Abuse Language Detection

Institute of Systems Science, National University of Singapore, Singapore 119615

ABSTRACT

Offensive Text is pervasive in social media. Individuals frequently take advantage of the perceived anonymity of computer-mediated communication, using this to engage in behavior that many of them would not consider in real life. Online communities, social media platforms, and technology companies have been investing heavily in ways to cope with offensive language in the form of text to prevent abusive behavior in social media.

The proliferation of, social media platforms resulted in a remarkable increase in user-generated content. These platforms have empowered users to create, share and exchange content for interacting and communicating with each other. However, these have also opened new avenues to cyberbullies and haters who can spread their negativity to a larger audience, often anonymously. Due to the pervasiveness and severity of this behavior, many automated approaches that employ natural language processing (NLP), machine learning and deep learning techniques have been proposed in the past. This survey offers an extensive overview of the state-of-the-art approaches proposed by the research community to identify offensive content. Based on our comprehensive literature survey, a categorization of different approaches and features employed by the researchers in the detection process are presented. This survey also incorporates the major challenges that require considerable research efforts in this domain. Finally, future research directions with an aim of developing robust abusive content detection system for social media are also discussed.

A specific popular form of online harassment is the use of abusive language. One abusive or toxic statement is being sent every 30 seconds across the globe. The use of abusive language on social media contributes to mental or emotional stress, with one in ten people developing such issues.These abusive Tweets and comments detection and deletion in social media is more important.

"The Online Criminal Harms Act will empower the Singapore government to issue directions to individuals, entities, online service providers and app stores requiring them to remove or block access to potentially criminal content."

1. INTRODUCTION

The last decade has seen a massive increase in user generated content in social media. While most people are interested in connecting with family and friends and in exchanging experiences, there is an increasing number of posts that cross the line from sharing negative opinions to becoming abusive. Since the data is too massive for manual filtering, automated methods to detect abusive language reliably are required. This has created a novel research area under the titles of abusive language detection, hate speech detection, flame or cyberbullying detection. While most of the work on abusive language detection has focused on English (Schmidt and Wiegand, 2017; Park and Fung, 2017; Lee et al., 2018), there is some work on other languages,and first attempts have also been made to develop methods that work across different languages (Fehn Unsvag and Gamb ack ¨ , 2018). Our interest also focuses on multilingual abusive language detection. However, before we engage in a full scale investigation of which methods work well across multiple languages, we need to know more about which factors have an effect on multilingual settings, including but not restricted to the compatibility of data and annotations, differences between languages, and topic effects. In the current paper, we focus on one languages, English

2. OVERVIEW

In this project, I have created and refined machine learning and Deep Learning models to detect Abuse posts in community help portal.

The goal of this project is to improve abusive language detection with a focus on implicit abuse, to develop a model using NLP techniques to accurately detect Abusive and Non-Abusive language Acquired data from Kaggle Competition for abuse language detection.

3. PROBLEM STATEMENT

Offensive Text and Image is pervasive in social media. Individuals frequently take advantage of the perceived anonymity of computer-mediated communication, using this to engage in behavior that many of them would not consider in real life.

Online communities, social media platforms, and technology companies have been investing heavily in ways to cope with offensive language in the form of text or images to prevent abusive behavior in social media.

4. DATASET

For our work, we chose data sets that were as similar as possible without creating a new, tightly controlled bilingual data set. For English, we used the publicly available Twitter hate speech data set created by Waseem and Hovy (2016).

Process the train data carefully as the data has emojis, English texts, some symbols, links etc. Also, note that the language detected often is not correct so don't rely blindly on it. Features like detected language of the text, total likes, total reports, and views along with text are also provided. These features were not included by me during the training process. Cleaned the data (remove emojis, punctuation etc.) Trim the data acc to text lengths.

Category	Count
Abuse	48602
Non-Abuse	363235

Fig. 1. Abuse Language Dataset.

5. DATA PREPROCESSING

Preprocessing of data is carried out before the model is built and the training process is executed. Following are the steps carried out during preprocessing.Initially the dataset is divided into training and validation sets.

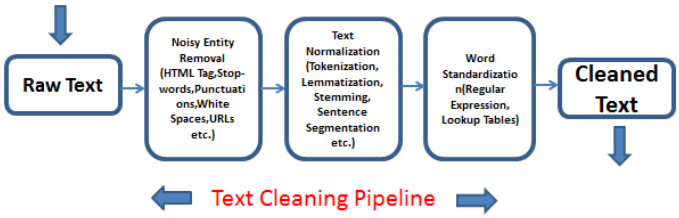

Fig. 2. Text Cleaning Pipeline

6. IMPLEMENTATION

We have developed pipeline is trained on English data to detect abusive language in tweets

Classifiers. For the first set of experiments, we use a range of classifiers including random forest, SVM, XGBoost,MLP. For the former three classifiers, we use scikitlearn (Pedregosa et al., 2011) Based on previous research on related tasks (Park and Fung, 2017; Badjatiya et al.,2017), we experiment with several promising architectures stand-alone and stacked respectively.For the scikit-learn classifiers, we optimized hyperparameters using grid search.

Features. We use simple character n-grams along with stemmed word n-grams and dependency parse-derived features. Stemming. Since we were unable to identify a good lemmatizer for English Twitter data, we decided to implement a stemmer for English data, stemmer (Majumder et al., 2007) is an unsupervised stemming algorithm that generates a minimum spanning tree of words based upon different string similarity distance metrics, cuts the hierarchy, and then stems a word by replacing the word with the centroid of the cluster the word belongs to. We use the YASS stemming method with a minor modification: While the YASS stemmer replaces all words that belong to a cluster with the cluster centroid, we replace all cluster members with the shortest member of the cluster. Stems are shorter than their morphologically related forms since affixes are not present. However, this step is not a departure from the YASS algorithm, instead it is an adaptation to increase the effectiveness of this algorithm on our particular domain. In addition, numbers, twitter handles, and URLs are removed from the data prior to stemming. The distance metrics used by Majumder et al. (2007) rely heavily on the suffixing nature of English inflectional morphology. While these distance metrics fall flat for non-suffix based inflectional morphology like irregular past tense (primarily ablaut grades in words like 'sleep', 'slept'), they produce less spurious stems compared to other common metrics like the Levenshtein distance. We use distance metric 4 from the YASS stemmer. Dependency parsing features. To extract dependency features for English, we use the Tweebo parser (Kong et al., 2014), which is designed to parse Twitter data and requires minimal preprocessing to obtain useful parses. Unlike English, parser pipeline (Bjorkelund et al. ¨ , 2010). However, in order to maximize the usefulness of Mate Preprocessing steps for parsing include: removing one or more hashtags or retweets after punctuation at the end of tweet as well as removing initial hashtags and retweets, removing the sign from any hashtag in the middle of the tweet, removing all emojis, and detaching punctuation from words. We also use a base list of abbreviations2 and add additional ones to ensure that these are kept during the tokenization process. We extract dependency triples consisting of (dependent, head, label) that occurred a minimum of five times as features. These features are Boolean valued, denoting their presence or absence in a tweet.

7. MODEL IMPLEMENTATION

MLP Classifier The multilayer perceptron (MLP) is a feed-forward artificial neural network model that maps input data sets to a set of appropriate outputs. An MLP consists of multiple layers and each layer is fully connected to the following one. The nodes of the layers are neurons with nonlinear activation functions, except for the nodes of the input layer. Between the input and the output layer there may be one or more nonlinear hidden layers.

- **hidden_layer_sizes :** With this parameter we can specify the number of layers and the number of nodes we want to have in the Neural Network Classifier. Each element in the tuple represents the number of nodes at the ith position, where i is the index of the tuple . . Thus, the length of the tuple indicates the total number of hidden layers in the neural network.
- **max_iter:** Indicates the number of epochs.
- **activation:** The activation function for the hidden layers.
- **solver:** This parameter specifies the algorithm for weight optimization over the nodes.

Why do we use XGBoost

We mainly use XGBoost because it offers many essential features that make it ideal for classification tasks. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples. Some of the reasons are given below:

- High performance: As mentioned above, XGBoost is optimized for speed and efficiency, making it appropriate for large datasets and real-time applications.
- Regularization methods: L1 (Lasso) and L2 (Ridge) regularisation terms are included in XGBoost to avoid overfitting and increase generalization.
- Handle missing data: Moreover, XGBoost can handle missing data automatically, minimizing the need for preprocessing and imputation.

Word2Vec Word2Vec creates vectors of the words that are distributed numerical representations of word features – these word features could comprise of words that represent the context of the individual words present in our vocabulary. Word embeddings eventually help in establishing the association of a word with another similar meaning word through the created vectors.

N-Grams usage

Normally we use some words more frequently than other. We can combine Zipf'slaw[16] with above statement and can state as : The occurrence of n most frequent word in text is proportional to 1/n. The general usage language consists on lot of words which are common. In some cases, when we are classifying same type of data then some words present in all groups. Those are not much useful while classifying the data. Generating frequency Profile:

- Data is modified by discarding numbers and punctuations, the necessary blank spaces are added to thedata.
- Then generate all possible n-grams (let's say 1 to 5) including the blanks aswell.
- Then fit the data into a hash table along with frequency such that each n-gram has its own count.
- Now sort these n-grams in descending order of occurrences then remove the count and store onlyngrams.

Usually the top n-grams are the common words we use frequently in the human language.Now comparing two texts using n-grams[13].After generating two frequency profiles one of each text,We measure the place of each n- gram in the profile with respect to another as shown in figure 2.

While classifying data into multiple groups we will take one group calculate sum of relative position of each n-grams of the text to that of the group. We will perform this on all group, then we will classify the text to a category that has the minimum sum. This is how the n-gram classification works.

8. BAG-OF-WORDS USAGE IN THE STUDY

In this model, a text is represented as the vector of its words, ignoringsentence structure and even the order of the words but keeping frequency. Frequency of each word is used as a feature to train the classifier[11].

- Allocate an integer id to every word in the text of the training set.
- For the text n, count the no. of existences of each word W and store it in X[n, m] as the value of feature m, m is the index of word W in the dictionary.

Here we have used CountVectorizer function available in the Sci-kit learn library of Python to convert the group of text documents to a sparse matrix representation [10]. There is an issue with the occurrence count that is longer the documents, higher the count values. We need to downscale the weights for words that occur in many text documents.This down scaling

is called TF-IDF which stands for "Term Frequency times Inverse Document Frequency".We use the TfidfTransformer() function of the Sci-kit learn library to produce the term frequencies from the matrix of token counts. After achieving the features, we train a classifier to predict the category of an article. Here we have implemented two different classifiers, Naive Bayes and Support Vector Machines, for predicting classes of documents in test data-set

Limitations of Bag-of-Words Approach:

Bag-of-words takes into account the existences of each word, neglecting the semantics and grammar of the natural language. Thus while dealing with Natural languages, we need to take into consideration, the usage of words, semantics and meaning of the sentence the words are a part of N-Grams is one such technique, where we vectorize not one but more than one words together, which convey much more information, than just the number of occurrences.

9. EXPERIMENTAL DESIGN

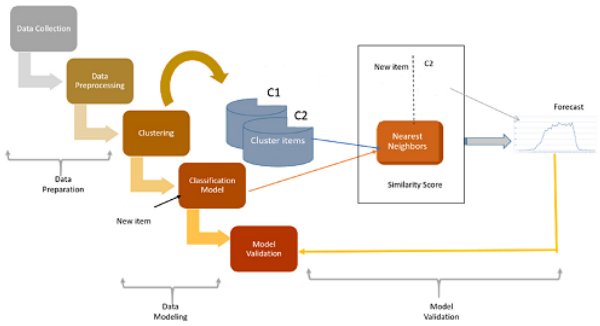

Fig. 3. Experimental design.

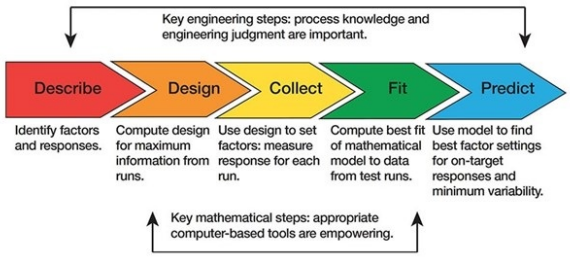

Fig. 4. Key Engineering Steps.

10. FEATURE SELECTION ACROSS LANGUAGES

The next set of experiments is concerned with the question of whether we can use the same features across languages, or if each language requires its own set of informative features. For these experiments, we decided to focus on MLP since they show good performance and similar trends across the languages in the comparison of classifiers above and since they train much faster than XGBoost. We add two additional feature types into the vectors: stems and dependency features.

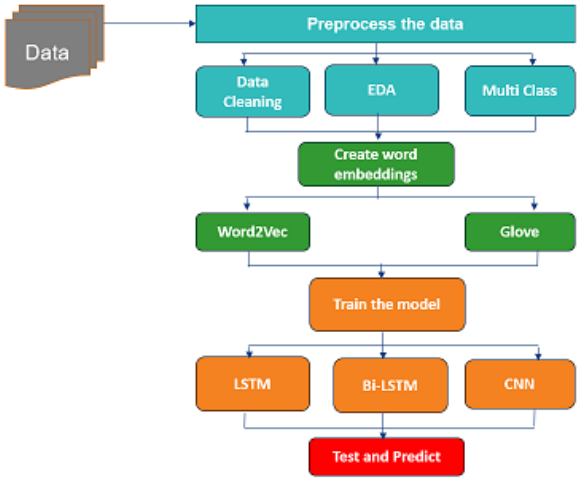

Fig. 5. Abuse Language Proposed system.

11. MODEL RESULT

Model	Accuracy
XGB	0.8551
MLP Word2Vec	0.8815
XGB Word2Vec	0.798
MLP	0.8608

Fig. 6. Abuse Language Model Accuracy.

- Among Machine Learning methods, MLP Word2Vec gave best accuracy 0.88%.
- XGB Ngram accuracy 0.85%
- MLP Ngram accuracy 0.85%
- XGB Word2Vec is low performer accuracy 0.79%
- Feature Extraction using IFIDF , Word2Vec techniques helped in improving overall accuracy.

12. FUTURE WORK

The Current project designed using only Content based feature. We can extend this project using User based, Activity based features

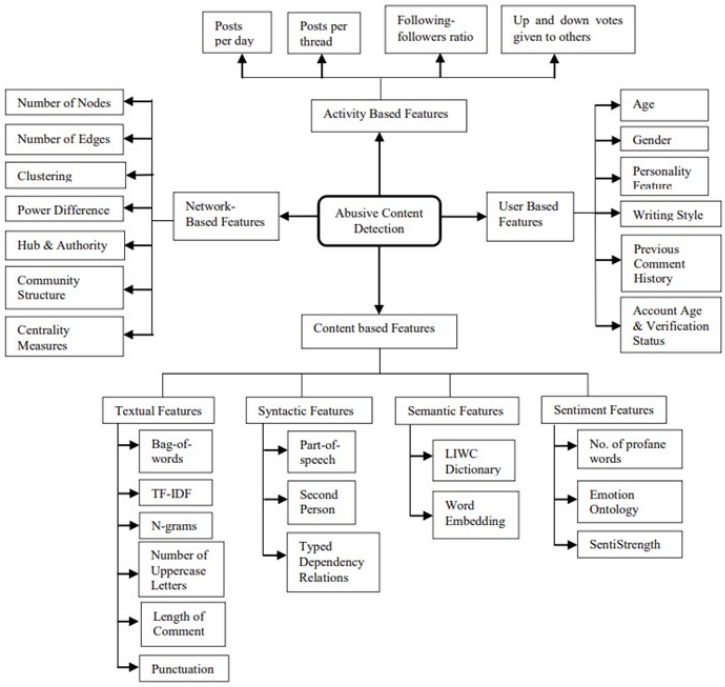

Fig. 7. Future Work .

13. MAJOR CHALLENGES/ISSUES INVOLVED

From the analysis illustrated above, abusive language detection in social media is a challenging task due to the unstructured, subjective, informal, and often misspelled nature of the textual content. This section covers the major challenges discovered in this task and the possible directions for future research.

- Lack of benchmark datasets: The main challenge for the abusive content detection is the dearth of benchmark datasets in this field. Presently, the researchers collect data from different social media platforms and annotate it using one of two methods—their own labeling effort or by using crowdsourcing services such as CrowdFlower and Amazon's Mechanical Turk. Due to unavailability of standard datasets, comparing different techniques is very difficult.

- Subjectivity involved: Another challenge is the myriad of forms and the lack of a clear, common definition of abusive behavior. Moreover, the notion of abusive or offensive is very subjective, and its degree varies considerably among people, making the labeling process more difficult.

- Sarcasm Detection: abusive comments may involve harassment without the use of profane words, satire, or irony, which are mainly difficult for machines to handle. For instance, the sentence, "You are as intelligent as Einstein" does not contain any profanity but may ironically be used to insult someone.

- Obfuscation: Simple keyword-based techniques fail in practice because of the intentional obfuscation of their text to evade keyword filters by the commenters. Strategies used by inventive users like symbol 280 Simrat Kaur et al. / Procedia Computer Science 189 (2021) 274–281 Author name / Procedia Computer Science 00 (2019) 000–000 7 substitution or false segmentations that still preserve the original semantics such as a$$hole or sh*t make these approaches ineffective [39].

- Context Sensitivity: Another challenge is to incorporate the context of the comment, especially in the case of threaded conversations. The potential for different interpretations of a word or sentence, if it is considered out of context, may impact the decision of the classifier. Based on the survey of numerous research articles, we give the following possible pertinent suggestions for future research: Most of the previous research in this field has performed binary classification only. So, there is a need to explore fine grained categories associated with abusive content such as insults, hate speech, threats etc. This fine-grained categorization will provide insight into different forms of inappropriate content and the degree to which they are alarming. For example,

14. CONCLUSION

This paper presented a comprehensive overview of the field by incorporating research articles spanning time of a decade and proposed taxonomies based on features and methods. The researchers have successfully applied methods from the machine-learning field, with Bag-of-Words (Bow) and N-grams being the most frequently used features in classification. Incorporation of more complex features i.e. from users' profile, activity statistics and social graph structure have also been shown to be effective in classification. Because of the high dimensionality and sparsity issues of the previous models, several recent works have employed distributed word representations also known as word embeddings. More recently, deep learning-based architectures have also shown encouraging results in this domain. To sum up, deeper linguistic features, analysis of demographic influences and clear annotation guidelines are required to differentiate between different types of abuse accurately. Despite the availability of large amounts of work, it remains difficult to judge the effectiveness and performance of various features and classifiers, mainly due to the use of different datasets by each researcher. To make comparative evaluation possible, clear annotation guidelines and a benchmark dataset are must. From the analysis, it is evident that abusive content detection is yet an open

area of interest for research fraternity, requiring more intelligent techniques to tackle the major challenges involved and thus making the online interaction a safer place for its users.

15. THE PROS AND CONS OF CHATGPT: UNVEILING THE POWER OF AI LANGUAGE MODELS

Are you currently considering the implementation of ChatGPT for your business? If so, you might be interested in learning more about the pros and cons of ChatGPT for your business.

ChatGPT, powered by OpenAI's advanced language model, is a revolutionary artificial intelligence technology that has transformed the way businesses interact with customers and streamline various processes. As the demand for personalized customer experiences and automation rises, ChatGPT has emerged as a game-changer.

This artificial intelligence technology has garnered significant attention in recent years, promising various benefits for companies in customer interactions and improving business process efficiency. However, like any other new technology, the use of ChatGPT also comes with pros and cons that need careful consideration before adoption.

We will explore some of the pros and cons related to using ChatGPT for your business, including its strengths and weaknesses. This can help you gain deeper insights before making this crucial decision.

What Benefits ChatGPT Can Do For Your Business? ChatGPT offers several benefits for businesses, making it a valuable tool in various aspects of their operations. Here are some of the key advantages:

- **Improved Customer Support** ChatGPT can handle customer queries and provide real-time support, offering quick and accurate responses. This enhances customer satisfaction and reduces the need for human intervention, allowing businesses to offer support 24/7.

  Data Insights ChatGPT can gather valuable data from customer interactions, providing businesses with insights into customer preferences, pain points, and commonly asked questions. This information can be used to improve products, services, and marketing strategies.

- **Personalized Interactions** ChatGPT can analyze user data and tailor responses to individual customers, providing more personalized interactions. This personalization helps in building stronger customer relationships and increasing customer loyalty.

- **Automation of Repetitive and Recurring Tasks** By integrating ChatGPT into the company's system, various routine tasks such as data and document management can be automated. This allows employees to concentrate on other more complex and value-added tasks

- **Content Creation** ChatGPT can be utilized to assist in content writing, such as articles, blogs, or product descriptions. This can enhance the marketing team's productivity and generate relevant and engaging content for the audience.

- **Candidate Screening in the Recruitment Process** In the recruitment process, ChatGPT can be employed to conduct initial screening of candidates based on specific qualification questions. This can simplify the selection process and save recruitment time.

- **Versatility** ChatGPT can be integrated into various platforms and applications, such as websites, mobile apps, and social media, making it a versatile tool for engaging with customers across multiple channels.

- **Improved Lead Generation and Conversion** ChatGPT can engage potential leads in personalized conversations, qualifying them based on predefined criteria. By capturing essential lead information and nurturing prospects through the sales funnel, ChatGPT enhances lead generation and conversion rates.

Advantages of ChatGPT

- **Cost-Effective Solution** Implementing ChatGPT as a customer support tool can lead to significant cost savings for businesses. Unlike traditional call centers that require a large team of agents to handle customer inquiries, ChatGPT automates the process, reducing the need for extensive human resources while maintaining high-quality interactions.

- **Scalability and Flexibility** ChatGPT's scalability and adaptability make it suitable for businesses of all sizes. It can handle numerous simultaneous conversations, ensuring that customer inquiries are addressed promptly, regardless of the volume. Moreover, its flexibility allows customization to match the brand's tone and personality.

- **Advanced Language Capability** ChatGPT possesses the ability to comprehend and generate human-like text at an advanced level, enabling it to interact with users naturally and intuitively.

- **Support for Multiple Languages** ChatGPT supports various languages, making it applicable in diverse environments and capable of summarizing information from multiple language sources.

- **Ability to Process Large Volumes of Data** With its capacity to analyze vast amounts of data and identify patterns across the entire dataset, ChatGPT can provide valuable business insights from large and diverse data sources.

Disadvantages of ChatGPT

- **Limitations in Understanding Context** While ChatGPT has made significant progress in understanding context, it can still produce irrelevant or nonsensical responses. The model's lack of understanding of complex contexts can lead to frustrating interactions with customers, potentially damaging the brand's reputation.

  Ethical Concerns and Bias AI language models like ChatGPT learn from vast datasets, which may contain biased information. As a result, the model may inadvertently produce biased or discriminatory responses, leading to ethical concerns and potential legal implications for businesses.

  Security and Privacy Risks Using ChatGPT to interact with customers may pose security and privacy risks, particularly when handling sensitive information. Businesses must ensure robust data encryption and security measures to safeguard customer data from unauthorized access. This technology can also be misused to spread false information, engage in spamming, or disseminate harmful or unethical content.

  Overreliance on Automation Relying heavily on ChatGPT for customer support may lead to a lack of human touch and empathy. Some customers prefer human interactions, especially in complex or emotionally charged situations, and may become dissatisfied with AI-driven responses.

Pros and Cons of ChatGPT: What are the perspectives?

Despite its advantages and capabilities, ChatGPT undoubtedly generates both pros and cons from various perspectives that shape its perception. The public's viewpoint regarding the use of ChatGPT can vary depending on factors such as their level of technological knowledge, experience, and understanding of the technology's strengths and limitations. Here are some perspectives that the public may hold regarding the pros and cons of ChatGPT:

- **Convenience and Accessibility** Many people appreciate the convenience of using ChatGPT to quickly obtain information, answers to questions, or assistance with tasks. Its availability on various platforms makes it easily accessible for users.
- **Suspicion and Distrust** People may feel skeptical or distrustful of ChatGPT due to their awareness of the risks of errors or information manipulation that could occur. They might feel more comfortable interacting with humans rather than machines.
- **AI Technology Dependence** There is a perspective that excessive reliance on AI technologies like ChatGPT might lead to reduced critical thinking and creativity in users, as they may come to depend heavily on AI-generated content.
- **Privacy and Data Security** Concerns about data privacy and the storage of conversations with ChatGPT arise, with some users being cautious about sharing sensitive information.
- **Education and Research** Academics and researchers often appreciate ChatGPT for its potential in various fields like natural language understanding, linguistics, and even creative writing. It provides a valuable tool for studying human language and exploring AI capabilities.
- **Depersonalization of Interactions** While ChatGPT offers quick responses, some individuals may miss the human touch and personalized interactions they have with human customer service representatives.
- **Job Displacement Concerns** Some express concerns about the potential impact of AI language models on job markets, especially in customer service and content writing fields. The automation of tasks previously handled by humans could lead to job displacement.

How to Know if ChatGPT is The Right Technology for Your Business?

Evaluating ChatGPT for your business requires careful consideration of several factors to ensure that it aligns with your specific needs and goals. Here are the steps to evaluate ChatGPT for your business:

- **Identify Business Objectives** Clearly identify the use cases and scenarios where you plan to use ChatGPT. Determine how it can add value to your business, such as improving customer support, automating tasks, or generating content.
- **What Are Advantages and Disadvantages?** Clearly understand the strengths and weaknesses of ChatGPT, such as communication efficiency, data analysis, or inaccuracies in context understanding. Compare these pros and cons with your business needs to determine if the expected benefits outweigh the risks and limitations of this technology. Familiarize yourself with the capabilities and limitations of ChatGPT. Know what types of tasks it can handle effectively and where human intervention might still be required.
- **Assess Use Cases** Identify potential use cases for ChatGPT within your organization. It could be for customer support, content generation, language translation, data analysis, or other relevant applications.
- **Data Privacy and Security** Ensure that the ChatGPT provider follows robust data privacy and security practices. Understand how user data will be handled and stored to protect sensitive information.

- **Integration with Other System or Application** Review the existing technological infrastructure in your company and consider how ChatGPT will be integrated into the system. Ensure that this technology can seamlessly function with the current systems in place and evaluate how easily ChatGPT can integrate into your existing systems and processes. Consider the level of training required for your team to use it effectively.

- **Data for Training ChatGPT** Ensure that you have sufficient and relevant training data to train ChatGPT effectively to work well within the context of your business. High-quality data will significantly impact this technology's ability to deliver accurate and relevant outcomes.

- **Long-Term Viability** Consider the long-term viability of ChatGPT. Assess its ability to adapt to evolving technologies and meet your business needs in the future.

- **Decision-making Process** Involve key stakeholders in the decision-making process based on the evaluation findings. Ensure that the final decision aligns with your business objectives. Consider whether the company has the resources and capabilities to manage and maintain ChatGPT effectively. This includes tasks such as training the model, monitoring performance, and addressing any potential issues that may arise.

- **Testing and Monitoring** Conduct limited trials with ChatGPT to understand its performance in real-world scenarios. Monitor the results and make necessary improvements before fully deploying it.

The Significance of ChatGPT in Today's Business

In today's fast-paced and digitally-driven business landscape, ChatGPT plays a pivotal role in enhancing customer experiences, improving operational efficiency, and gaining a competitive edge. With its ability to handle large volumes of inquiries and provide real-time assistance, ChatGPT empowers businesses to meet customer expectations and adapt to evolving market demands.

The Combination of ChatGPT and Automation

The combination of ChatGPT and automation has opened new doors in how businesses interact with customers and manage processes more efficiently. By leveraging the artificial intelligence of ChatGPT and the power of automation, companies can achieve unparalleled levels of customer service.

ChatGPT functions as a responsive virtual assistant, providing quick and accurate answers to customer inquiries and delivering a more personalized experience by recognizing user preferences. On the other hand, automation takes over routine tasks, optimizing business processes and enhancing operational efficiency. This allows employees to focus more on tasks that require creative and analytical thinking.

Moreover, this combination unlocks the potential for faster data analysis, generating valuable business insights, and facilitating smarter decision-making. However, it is crucial to carefully consider how to integrate these two technologies wisely, taking into account business needs, data security, and privacy concerns to achieve optimal results for both the company and its customers.

16. AUTHOR

Thota Siva Krishna , e0943696@u.nus.edu

17. REFERENCES

[1] Sourander, A., Klomek, A. B., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen M., & Helenius, H, (2010) "Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study", Archives of general psychiatry, 67(7), 720-728.

[2] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, L. Edwards (2009) "Detection of harassment on Web 2.0" In Proceedings of Content Analysis in the WEB 2.0 (CAW 2.0) Workshop at WWW, Madrid, April 20-24, 2009.

[3] K. Dinakar, R. Reichart, H. Lieberman (2011) "Modelling the Detection of Textual Cyberbullying." In ICWSM 2011, Spain, July 17-21.

[4] Chen,Y., Zhu, S., Zhou, Y., Xu, H. (2011) "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety." In: Symposium on Usable Privacy and Security. Pittsburgh, USA.

[5] Sood, S. O.; Churchill, E. F.; and Antin, J. (2012) "Automatic identification of personal insults on social news sites." Journal of the American Society for Information Science and Technology, 270-285.

[6] Munezero, M., Montero, C.S., Kakkonen, T., Sutinen, E., Mozgovoy, M. and Klyuev, V. (2014) "Automatic detection of antisocial behaviour in texts." Informatica, 38(1), p.3.

[7] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.(2016) "Abusive Language Detection in Online User Content." In Proceedings of the 25th International Conference on World Wide Web, pp. 145-153.

[8] Dadvar, Maral, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong (2013) "Improving cyberbullying detection with user context." In European Conference on Information Retrieval, pp. 693-696. Springer Berlin Heidelberg.

[9] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali (2017) "Mean Birds: Detecting Aggression and Simrat Kaur et al. / Procedia Computer Science 189 (2021) 274–281 281 8 Author name / Procedia Computer Science 00 (2019) 000–000 Bullying on Twitter." In Proceedings of the 2017 ACM on Web Science Conference (pp. 13-22). ACM.

[10] Chen, H., Mckeever, S., Delany, S. J. (2017) "Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media." In Advances in Computational Intelligence Systems (pp. 187-205). Springer International Publishing.

[11] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber (2017) "Automated Hate Speech Detection and the Problem of Offensive Language." In Proc. of ICWSM.

[12] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana (2016) "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network." Computers in Human Behavior, 63, 433-443.

[13] Xu, Z., Zhu, S. (2010) "Filtering offensive language in online communities using grammatical relations." In Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (pp. 1-10).

[14] Justo, R., Corcoran, T., Lukin, S. M., Walker, M., Torres, M. I. (2014) "Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web." Knowledge-Based Systems, 69, 124-133.

[15] Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. (2015) "Antisocial behavior in online discussion communities." In Proceedings of ICWSM. Menlo Park, California: AAAI Press.

[16] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N. (2015) "Hate speech detection with comment embeddings." In Proceedings of the 24th international conference on world wide web (pp. 29-30). ACM.

[17] Zhao, R., Zhou, A., Mao, K. (2016) "Automatic detection of cyberbullying on social networks based on bullying features." In Proceedings of the 17th international conference on distributed computing and networking (p. 43). ACM.

[18] E. Cambria, D. Olsher, D. Rajagopa (2014) "Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis." In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI Press.

[19] Dadvar, M., de Jong, F.M.G., R.J.F., Ordelman, Trieschnigg, D.(2012) "Improved Cyberbullying Detection Using Gender Information." In: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop, pp. 23-26., Belgium.

[20] Balakrishnan, V. (2015) "Cyberbullying among young adults in Malaysia: the roles of gender, age and internet frequency." Computers in Human Behavior, 46, 149e157.

[21] Q. Huang, V. K. Singh and P. K. Atrey (2014) "Cyber bullying detection using social and textual features." The 3rd ACM MM Workshop on Socially Aware Multimedia (IWSAM14), Orlando, FL, USA.

[22] A. Kontostathis, L. Edwards, and A. Leatherman (2009) "ChatCoder: Toward the tracking and categorization of Internet predators." In Proceedings of the Text Mining Workshop.

[23] K. Reynolds , A. Kontostathis , L. Edwards (2004) "Using Machine Learning to Detect Cyberbullying." Proceedings of the 10th International Conference on Machine Learning and Applications and Workshops, p.241-244, December 18-21.

[24] Waseem, Z., Hovy, D. (2016) "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter" In SRW@ HLT-NAACL (pp. 88-93).

[25] Xiang, G., Hong, J., Rosé, C. P.(2012) "Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus." Proceedings of the 21st ACM Conference on Information and Knowledge Management, Sheraton, Maui Hawaii, 2012.

[26] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[27] Nahar, V., Al-Maskari, S., Li, X., Pang, C (2014). "Semi-supervised Learning for Cyberbullying Detection in Social Networks." In ADC, pp. 160-171, 2014.

[28] Di Capua, M., Di Nardo, E., Petrosino, A. (2016) "Unsupervised cyber bullying detection in social networks." In Pattern Recognition (ICPR), 2016 23rd International Conference on (pp. 432-437). IEEE.

[29] Marzieh Mozafari, Reza Farahbakhsh, Noel Crespi.(2019) "A BERT-based transfer learning approach for hate speech detection in online social media. Complex Networks" 8th International Conference on Complex Networks and their Applications, pp.928-940.

[30] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A. (2017) "Hate is not Binary: Studying Abusive Behavior of GamerGate on Twitter." ACM Hypertext, 2017.

[31] Sood, S., Antin, J., Churchill, E. (2012) "Profanity use in online communities." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1481-1490). ACM.

[32] Razavi, A. H., Inkpen, D., Uritsky, S., Matwin, S (2010) "Offensive language detection using multi-level classification." In Canadian Conference on Artificial Intelligence, pp. 16-27, Springer Berlin Heidelberg.

[33] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012) "Common sense reasoning for detection, prevention, and mitigation of cyberbullying." ACM Trans. Interact. Intell. Syst. 2, 3, Article 18(September 2012).

[34] Dadvar, M., Trieschnigg, D. and de Jong, F., May (2014) "Experts and machines against bullies: A hybrid approach to detect cyberbullies." In Canadian Conference on Artificial Intelligence, pp. 275-281, Springer International Publishing.

[35] Badjatiya, P., Gupta, S., Gupta, M., Varma, V. (2017) "Deep learning for hate speech detection in tweets". In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759-760). International World Wide Web Conferences Steering Committee.