Annonymization of multimodal data

Muhammad Ammar Ahmed
Ludwig Maximilian University
Munich, Germany
Ammar.Ahmed@campus.lmu.de

Vladana Djakovic
Ludwig Maximilian University
Munich, Germany
Djakovic.Vladana@campus.lmu.de

***Abstract*—Nowadays, we witness the informatics and technical revolution permeating every sphere of our life. New technologies are enabling us to have more volume of data of varying types. Due to their open-source nature, these datasets are accessible to the public. These datasets are being leveraged by Artificial Intelligence (AI) models to improve everyday life. We notice the broader and more diverse implementation of AI techniques, from object and text detection to creating an image from text. Multiple algorithms are being developed to analyze images, text, audio, etc.**

Considering the fact that the tasks have become more diverse, they require different data types, for example, the dataset needs to contain pairs of images and text to generate an image from text. The data used to solve this and similar tasks are called multimodal data. Maintaining the privacy of these data sets is challenging and requires special techniques for each data type. Different techniques of anonymization should be applied in order to preserve the privacy of the whole dataset. This project addresses the anonymization of a dataset with tabular and textual data. We implement these techniques and point out their advantages.

***Key Words*—multimodal data, privacy, anonymization, name entity recognition, k-anonymity**

***Source code*—https://github.com/ammarlam10/DS23_multimodal_anonymization**

I. Introduction

Due to the technological revolution, previously sensitive data, such as data in hospitals, government offices, banks, etc., have been digitized and available for research purposes. This creates new challenges for researchers and exposes almost every person worldwide to fraud and identity theft. On top, Open Source Initiatives are creating more pressure to make each data set public, moreover protecting sensitive data is imperative.

Different data types cover different information, images, audio, and time changes. These data types are inspired by people's sense and ability to perceive the world around them [1]. Modality describes how something occurs or is perceived, and a research issue is classified as multimodal when it encompasses various modalities [2]. The most commonly used modalities are images, text, tabular data, time series, video and audio. Each case requires a different approach and dataset combination of modalities.

For example, if we are looking at a patient medical chart, it includes tabular data, images and text. Each of these modalities contains patient information and should be anonymized properly. Patient name, id, and gender should be removed from images before it is publicly available. Medical text can not contain personal information; the same stands for tabular data.

Another example is bank data, which includes a combination of time series, tabular and text data. Again each modality requires a different approach to maintain the privacy of customers. For each combination of types, multiple application cases can be created.

II. Anonnymization

In order to protect the privacy of the subject, the process of anonymization of data should be done. It requires removing direct and indirect personal identifiers based on which person could be identified. Direct Ids, such as a name, address, postcode, telephone number, etc. identify a person directly. All other information pieces linked to each other, which can directly determine the person, are called indirect identifiers. Those include a place of work, job title, salary, postcode or even a particular diagnosis or condition. [3]

A. Anonnymization of tabular data

In statistics, tabular data refers to the data organized in a table [5]. Most of the time, each row usually represents one subject, and columns store personal information associated with a subject. If the dataset contains personal information, anonymization should be performed following General Data Protection Regulation (GDPR) which can be achieved in different ways [4]. A few examples are masking, encryption, generalization etc.

To determine how data should be anonymized so that the subject cannot be either directly or indirectly identified, the concept of k-anonymity has been developed. We say that tabular data satisfy k-anonymity when the information for each person in the table cannot be differentiated among at least from k-1 others persons [6]. The attributes that, in combination, can uniquely identify individuals, such as birth date and gender, are called quasi-identifiers. Any quasi-identifier present in the released table must appear in at least k records. To achieve more privacy and anonymize data, the technique of l-diversity was introduced. L-diversity aims to ensure that each set of rows with identical quasi-identifiers has at least l distinct values [7].

In this work, we will focus on achieving the k-anonymity of tabular data. To achieve k-anonymity, multiple algorithms

have been created. Some of them are Datafly and Topdown Greedy, but our focus would be Basic Mondrian.
Mondrian was introduced by Kristen LeFevre [8] and is a Top-down greedy data algorithm for anonymizing relational datasets. Originally,this algorithm was made for numerical data, but nowadays, it can be used for text data. The idea is to split categorical attributes through generalization hierarchies, which means that categorical values would be transformed into general values and ranges. By doing this, we are creating a hierarchy among data and anonymizing it.

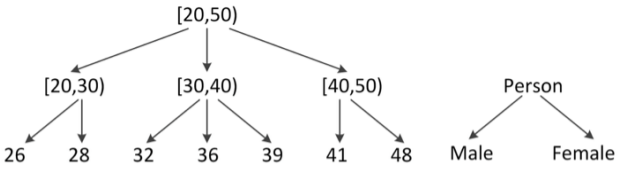

Fig. 1. Generalization hierarchies [9]

B. Anonnymization of text data

Natural Language Processing (NLP) is one of the most researched fields of Deep Learning. After the appearance of Attention (Vaswani et al.) in 2018, it has become an area of constant development. Expanding its application to different languages, tasks, and texts requires more training data than previous problems. The model development consists of two parts: pre-training the model and fine-tuning it.
The pre-training data is usually some large corpus publically available. The second part of model training is called fine-tuning. For this part, a task-specific dataset is given to the model, which is improved on that dataset. The fine-tuning dataset can be smaller in size and contain much private information.
Expanding the domain from where datasets were obtained, the problem of detecting Personally Identifiable Information (PII) has attracted more attention [10]. Various approaches in NLP have been created to detect different types of PII entities in free text.

- If entities share the pattern, leveraging the **Regular Expressions** and context of words could be used.
- For a finite list of options, **blacklists** could be used.
- **Rule-based recognizers** are used for entities which can be identified using specific logic.
- **Named Entity Recognition (NER)** is used for entities that require natural language understanding of the input.

NER is a process in which a sentence or chunk of text is parsed to find entities that can be put under categories like names, organizations, locations, quantities, monetary values, percentages, etc. [11]. In this work, we use combination of Regular expression and Named Entity Recognition to identify and anonymize data a method also used in Presidio. [12]
Presidio, created by Microsoft, helps to ensure that sensitive data is appropriately managed and governed. It provides fast identification and anonymization modules for private entities in text and images such as credit card numbers, names, locations, social security numbers, bitcoin wallets, US phone numbers, and financial data. This algorithm is open-source and available to anyone.
Presidio algorithm consists of the following 8 steps [10]:

1) Getting a request for anonymization from the user;
2) Passing request to Presidio-Analyzer for PII entities identification;
3) Extracting NLP features (lemmas, named entities, keywords, part-of-speech etc.) to be used by the various recognizers;
4) Fetching all PII recognizers (predefined + custom from the Recognizer Store service);
5) Running all recognizers;
6) Aggregatinge results;
7) Passing to Presidio-Anonymizer for de-identification,
8) Returning de-identified text to the calle.r

Fig. 2. Presidio workflow

The model authors have focused on improving NER rates for personal names, locations and organizations using different models and datasets in their work. For training purposes, they started with a labelled dataset (e.g., OntoNotes or CoNLL-2003) and processed it to extract templates.
For example, the sentence 'Hello Mark!' was first manually labelled, where the word Mark was replaced with the label [PERSON]. The result was, 'Hello [PERSON]!' Afterwards, more examples were labelled during the generation process. As the authors stated, this process was time-consuming and faced a few problems since the model needs to understand the context. Some examples they pointed out were nationalities, gender, personal names in institution names, and real-life knowledge, which were tackled separately and introduced more labels, for example, nationalities. However, some of these problems were left open to debate.

Fig. 3. NER example [10]

After pre-processing, they evaluated Conditional Random Fields, spaCy-based, and Flair-based models.

1) **spaCy** is production grade NLP library for tokenization, part-of-speech tagging entity extraction, etc. It contains a CNN model

2) **Flair** is built on top of PyTorch and features both special embedding techniques (called Flair embeddings) and prediction models. It can easily integrate with other embedding models like BERT and ELMo
3) **Conditional Random Fields (CRFs)** are a class of methods for sequence tagging. These discriminative graphical models learn the dependencies between predictions and are a natural fit for Named Entity Recognition tasks.

The metric used to compare these models is the F2 metric(recall is more important than precision) of the PII/no-PII binary decision. The results showed that Flair-based models were superior comparing the others.

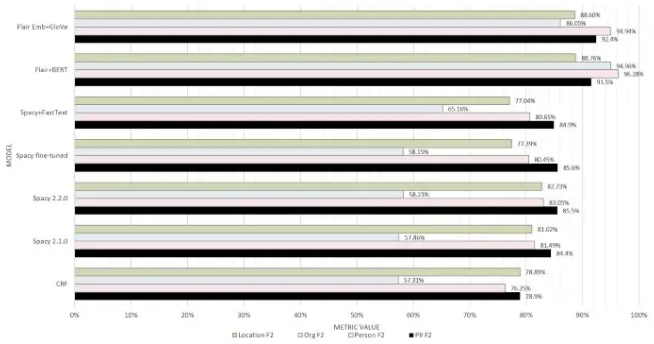

Fig. 4. Comparison of models [10]

III. Practical Implementation

1) Dataset: The inspiration for this project is one of the highly sensitive datasets- medical. The dataset comes from the LMU Clinic and is used for research purposes of employees. Due to privacy reasons, it could not be used in this project, but it consists of two parts:

- **tabular part** - patient personal information such as first name, surname, patient id, insurance number, street etc.
- **text part** - written report from a doctor.

To imitate that data format, different datasets were combined:

- **tabular part** - We have used the publicly available dataset *BPDEmployee.csv*[1], which was concatinated with radnomly generated personal information[2]. Adding personal information was necessary to fill in identifiers that would be in the original dataset.
- **text part** - Here we have used the open source dataset "Enron Email Dataset"[3]. This dataset contains emails from 150 company employees, mostly senior management of Enron. Due to limited capacities, we randomly sampled 1000 emails from the whole dataset. Those emails have a few PIIs, like names, addresses, phone numbers, emails, etc., which need to be anonymized.

[1]https://data.world/baltimore/baltimore-police-employees/workspace/file?filename=BPD_Employee_File_As_Of__06_09_2016.csv

[2]https://www.mockaroo.com

[3]https://www.cs.cmu.edu/ enron/

2) Preprocessing dataset:

- **tabular part** - The dataset *BPDEmployee.csv* contains information about active employees of the Baltimore Police Department. The information included is their rank, assignment, race, gender, age and years of service. To complete the dataset, we randomly generated personal information like name, surname, email. Two datasets were concatenated based on gender. This led to the creation of the dataset with a similar ratio of male and female officers, which can be a problem with anonymization.
- **text part** - For training purposes, training dataset has to be created manually. Each email from the training set needed to be manually tagged. In these emails, we have tagged personal information such as email, address, names, locations, phone number. The process of tagging the whole dataset would be time-consuming, so the random sample of 100 examples was extracted and processed. Additional ones are randomly generated from them to have more training dataset.

A. Anonnymization of tabular data

In order to anonymize the tabular part of the dataset for this project, the Mondrian algorithm, based on [13]. In this part of the dataset, we had several columns based on which dataset needed to be anonymized. The columns also had a different number of unique values, some of which are:

- AGE with 45 different values
- RACE/ETHNICITY with 10 different values
- YEARS_IN_SERVICE with 41 different values

Anonymization of this dataset using the Mondrian algorithm meant that values of indirect identifiers would be grouped in specific ranges based on their closeness.

For example, when we are talking about race and ethnicity, police employees that are of the same race and ethnicity and share other similar identities should be grouped. That means the officers that have the same assignment, degree, and title and share the same race and ethnicity should be grouped with a new variable count. This variable represents the number of police officers with the same identifiers. Similarly, the dataset is anonymized for each other category.

Additionally, the dataset can be further anonymized if multiple entities are combined, which is one of the suggestions for how anonymization can be improved. The other improvements include achieving diversity among the identifiers.

Since the dataset had personal information such as name and email address, that information was masked, and only the first letter of name and surname were kept. This could be a problem if the dataset is small and persons could be identified with prior knowledge. The solution of whether to delete these identities depends on the dataset's task and purpose.

This work focuses more on distinguishing the influence of the range of k on anonymization. The same anonymization has been performed for three values, $k = [2, 3, 4]$. The aim is to see how much the anonymization is improved when the

higher value of k has been selected. Pictures below show how the dataset's structure when anonymizing regarding race and ethnicity have changed with values of k.

POLICE OFFICER	NORTHERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	34.0	1
POLICE OFFICER	NORTHWESTERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	27.0	1
POLICE OFFICER	NORTHWESTERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	35.0	1
POLICE OFFICER	NORTHWESTERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	36.0	1
POLICE OFFICER	NORTHWESTERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	40.0	1

Fig. 5. Anonmyzed dataset for k=2

POLICE OFFICER	NORTHERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	7.0	2
POLICE OFFICER	NORTHERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	15.0	1
POLICE OFFICER	NORTHWESTERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	5.0	1
POLICE OFFICER	NORTHWESTERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	9.0	2
POLICE OFFICER	NORTHWESTERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	18.0	1

Fig. 6. Anonomyzed dataset for k=3

POLICE OFFICER	SOUTHERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	AFRICAN AMERICAN	4
POLICE OFFICER	WESTERN DISTRICT,SOUTHWESTERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	AFRICAN AMERICAN	7
POLICE OFFICER	NORTHERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	AFRICAN AMERICAN	2
POLICE OFFICER	NORTHERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	OTHER HISPANIC	1
POLICE OFFICER	NORTHWESTERN DISTRICT	HIGH SCHOOL GRADUATE	NO	FEMALE	BPD	AFRICAN AMERICAN	4

Fig. 7. Anonomyzed dataset for k=4

From these results, it can be noticed that for k=2 the majority of the dataset is not anonymized and that individuals can still be distinguished based on identifiers. However, as k has been increased, the identification of persons has become more challenging. For k=4, it is almost impossible to identify exact persons from the dataset. It needs to be noticed that this dataset is highly imbalanced regarding gender, and it consists of more male officers, making identifying the latter easier.

B. Anonnymization of text data

We have implemented a reduced version of the Presidio algorithm for this project. The official GitHub of the Presidio algorithm [12] was used as a guidance and reference for functions. Our implementation consists of three steps:

1) As described in pre-processing. We have chosen the random sample of the Enron dataset. A part was manually labelled, and an additional one was generated from that sample.
2) Fine-tune 2 NER models from the SpaCy package:
   - en_core_web_sm - a small English pipeline trained on written web text (blogs, news, comments) that includes vocabulary, syntax and entities [14].
   - en_core_web_trf - English transformer pipeline (Roberta-base). Components: transformer, tagger, parser, ner, attribute_ruler, lemmatizer [14].
3) Evaluating both models and choosing one with a better F1. That model was en_core_web_trf.

Once the NER model is fine-tuned on the dataset, it can extract entities from Enron emails. For this dataset, it was noticed that the following entities needed to be mapped: 'PERSON', 'LOCATION', 'GPE', 'ORGANIZATION', 'DATE_TIME', and 'NRP'. The performance of the NER model was measured for each identity, and the results are shown in the next chapter. Besides these entities, this dataset contained a lot of personal emails and passwords, which should also be anonymized. Here pre-train regex was used. A regular expression (regex) is a text string that can create patterns with whom it is easier to match, locate and manage text [15]. The model was now able to recognize patterns and characters in order to detect emails and passwords.

Finally, all the extracted entities are anonymized and replaced by placeholder values. The example of input and output emails are shown below. In the input email, information like name, date, and organization can be seen, but after anonymizing, this information is covered by placeholder value.

```
 -----Original Message-----
From:   "Eva Pao" <<EMAIL>>@ENRON [mailto:IMCEANOTES-
+22Eva+20Pao+22+20+3Cepao+40mba2002+2Ehbs+2Eedu+3E+<EMAIL>]
Sent:   Wednesday, June 20, 2001 2:31 PM
To:     John Arnold
Subject:        soccer gear

for tonight's game, i've got a manchester united jersey.  will people jeer
at me because it is english or because it looks like the trinidad jersey?

kid
```

Fig. 8. Example of email

```
 From :          " Eva Pao " < < EMAIL>>@ENRON [ mailto : IMCEANOTES-+22Eva+20Pao+22 + 20 +
3Cepao+40mba2002 + 2Ehbs+2Eedu+3E+<EMAIL > ]
 Sent :          <DATE> <DATE> <DATE> <DATE> <DATE> <DATE> <TIME> <TIME>
 To :    <PERSON> <PERSON>
 Subject :       soccer gear

 for <TIME> 's game , i 've got a <ORG> <ORG> jersey .   will people jeer
 at me because it is english or because it looks like the <ORG> jersey ?

 kid
```

Fig. 9. Example of anonymized email

Further examples of anonymization of emails are shown in Appendix. Additionally, the model was adapted to anonymize arbitrary text given by the user and return the anonymized result. One example is:

```
'My name is Ammar Ahmed , and I m studying in LMU . Ill make the transaction tomorrow'
```

Fig. 10. Example of unanonymized text

```
'My name is <PERSON> <PERSON> , and I m studying in <ORG> . Ill make the transaction <DATE>'
```

Fig. 11. Example of anonymized text

IV. Analysis of results

When talking about the performance of these models, it is important to include the subjectiveness of research. Depending on the task, different standards may apply. In our

project, we have measured the performance based on selected entities.

A. Anonnymization of tabular data

Achieving K-anonymity of the tabular part of the dataset was highly dependent on which entities were selected and how large k is. Of interest was the dataset size when different $k = [2, 3, 4]$ has been selected. The original dataset had 809 persons and was highly imbalanced regarding sex. The dimensionality results concerning k and identifiers are shown in the table below.

TABLE I
DIMENSINOS OF DATASET WITH DIFFERENT K VALUES

k	AGE	YEAR IN SERVICE	RACE/ETHNICITY
2	739	739	721
3	724	724	698
4	404	404	323

From these results, the dataset size drastically reduces by increasing the value of k. However, it can be noticed that identifiers "AGE" and "YEAR IN SERVICE" are highly correlated and have different similar values. This explains why anonymization is similar when we compare them. The identifier "RACE/ETHNICITY" has fewer different values, and anonymization is achieved more straightforwardly. However, it can be noticed that when k=4, the size of the dataset is less than half of the original one. For further research, combining both identifiers could lead to even better results.

B. Anonnymization of text data

As NER is aiming to select all required entities from the text, the measurement of its performance are the standard ones:

- Precision is a fraction of relevant instances among the retrieved instances.
- Recall is a fraction of relevant instances that were retrieved.

In our case, we have measured how well NER could recognize each entity in the text, and the results are shown in the figure below.

Entity	Precision	Recall	Number of samples
ORG	79.28%	60.88%	4696
GPE	34.02%	65.66%	530
PERSON	92.53%	86.38%	4032
DATE	89.49%	83.59%	7417
PII	85.82%	86.04%	16675

PII F measure: 86.01%

Fig. 12. Results NER

Here, it can be seen that NER had the highest precision when tagging the person's name and surname, almost 93%, and the highest recall, 84%. On the other hand, the model could not detect the Geopolitical Entity (GPE) having the lowest precision and recall. Such an outcome is also possible since the number of samples for this entity was significantly lower than for other entities. For entities that were given by regex, the model was already trained to recognize them, so no performance metrics were calculated on them.

V. CONCLUSION

To sum up, the anonymization of multimodal data brings various challenges and can be subjective. When talking about tabular data, anonymizing is highly dependent on the task and the number of different values of identifiers. Usually, for research purposes, tabular data is pseudonymized, and it can be continuously performed.

On the other hand, anonymizing text data is still an ongoing challenge. As we have seen, the labelling of text is time-consuming, and models require plenty of labels in order to perform well. An additional problem is a language itself. These days NLP is spreading to different languages, where despite of insufficient labelled text, the language itself can create problems, as it is in German, where all nouns have capital letters, not just names. All these challenges make anonymizing text data an essential research point.

REFERENCES

[1] O. E. C. D. S. Directorate, OECD Glossary of statistical terms - data type definition. [Online]. Available: https://stats.oecd.org/glossary/detail.asp?ID=3016. [Accessed: 29-Jan-2023].

[2] P. Mehta, "Multimodal Deep Learning," Medium, Jun. 13, 2020. https://towardsdatascience.com/multimodal-deep-learning-ce7d1d994f4

[3] Ucl (2019) Anonymisation and Pseudonymisation, Data Protection. Available at: https://www.ucl.ac.uk/data-protection/guidance-staff-students-and-researchers/practical-data-protection-guidance-notices/anonymisation-and (Accessed: January 30, 2023).

[4] "What Is Data Anonymization? Definition and FAQs — HEAVY.AI." Www.heavy.ai, www.heavy.ai/technical-glossary/data-anonymization. Accessed 4 Feb. 2023.

[5] Zach (2022) What is tabular data? (definition amp; example), Statology. Available at: https://www.statology.org/tabular-data/ (Accessed: January 30, 2023).

[6] Shmatikov, Vitaly. Data Privacy., Stanford University

[7] Machanavajjhala, Ashwin, et al. Diversity: Privacy beyond KAnonymity.

[8] K. LeFevre, D. J. DeWitt, R. Ramakrishnan. Mondrian Multidimensional K-Anonymity ICDE '06: Proceedings of the 22nd International Conference on Data Engineering, IEEE Computer Society, 2006, 25

[9] Gong, Qiyuan. "Basic Mondrian." GitHub, 29 June 2022, github.com/qiyuangong/Basic_Mondrian.

[10] Mendels, Omri. "NLP Approaches to Data Anonymization." Medium, 8 Jan. 2020, towardsdatascience.com/nlp-approaches-to-data-anonymization-1fb5bde6b929. Accessed 4 Feb. 2023. .

[11] "Named-Entity Recognition." DeepAI, 17 May 2019, deepai.org/machinelearningglossaryand-terms/namedentityrecognition. Accessed 4 Feb. 2023.

[12] "Microsoft Presidio." Microsoft.github.io, microsoft.github.io/presidio/.

[13] Fujita, Taisuke. "AnonyPy." GitHub, 24 Jan. 2023, github.com/glassonion1/anonypy. Accessed 4 Feb. 2023.

[14] "Trained Models & Pipelines · SpaCy Models Documentation." Trained Models & Pipelines, spacy.io/models. Accessed 4 Feb. 2023.

[15] "What Is a Regex (Regular Expression)?" Www.computerhope.com, www.computerhope.com/jargon/r/regex.htm.

VI. Appendix

```
Also ,   is the trader <PERSON> <PERSON> ?   Please give me the name of a contact w/
 phone number as well as complete address for <ORG> .

 Thanks !

 <PERSON> <PERSON>
 <ORG> <ORG> <ORG> <ORG>
 Legal Department
 1400 Smith Street , EB 3885
 <GPE> , <GPE> 77002
 < EMAIL >
 Phone 713 - 853 - 7658
 Fax   713 - 646 - 3490
------------------
Also,  is the trader Paul Lucci?  Please give me the name of a contact w/
phone number as well as complete address for Cenex.

Thanks!

Debra Perlingiere
Enron North America Corp.
Legal Department
1400 Smith Street, EB 3885
Houston, Texas 77002
<EMAIL>
Phone 713-853-7658
Fax  713-646-3490
```

Fig. 13. Example 1.

```
Please be advised of the name change . ? My last name is now <PERSON> . Please
 make the necessary changes so that I may still receive my e - mail . ? Thanks &
 if you have any questions please call me .

 <PERSON> <PERSON>

 <ORG> <ORG> <ORG>

 1200 Lathrop Street

 <GPE> , <GPE> ? 77020

 Phone : 713 - 671 - 8200 X116

 Fax : 713 - 671 - 8225 or 8229

 < EMAIL >
------------------
Please be advised of the name change.? My last name is now Hernandez. Please
make the necessary changes so that I may still receive my e-mail.? Thanks &
if you have any questions please call me.

Dora Hernandez

UnitedDC, Inc.

1200 Lathrop Street

Houston, TX? 77020

Phone: 713-671-8200 X116

Fax: 713-671-8225 or 8229

<EMAIL>
```

Fig. 14. Example 2.

```
<PERSON> ,

 Could you move the PMA adj . for Power ( 423 , 818 ) to the power roll ? <PERSON> needs to see this in the power roll , so th
at she can net it against her outstanding variances .

 Thanks !

 <PERSON> <PERSON>
 <ORG> <ORG> <ORG>
 713 - 853 - 5944
 < EMAIL >
------------------
Errol,

Could you move the PMA adj. for Power (423, 818) to the power roll? Tracy needs to see this in the power roll, so that she can
net it against her outstanding variances.

Thanks!

Shannon McPearson
Enron North America
713-853-5944
<EMAIL>
```

Fig. 15. Example 3.