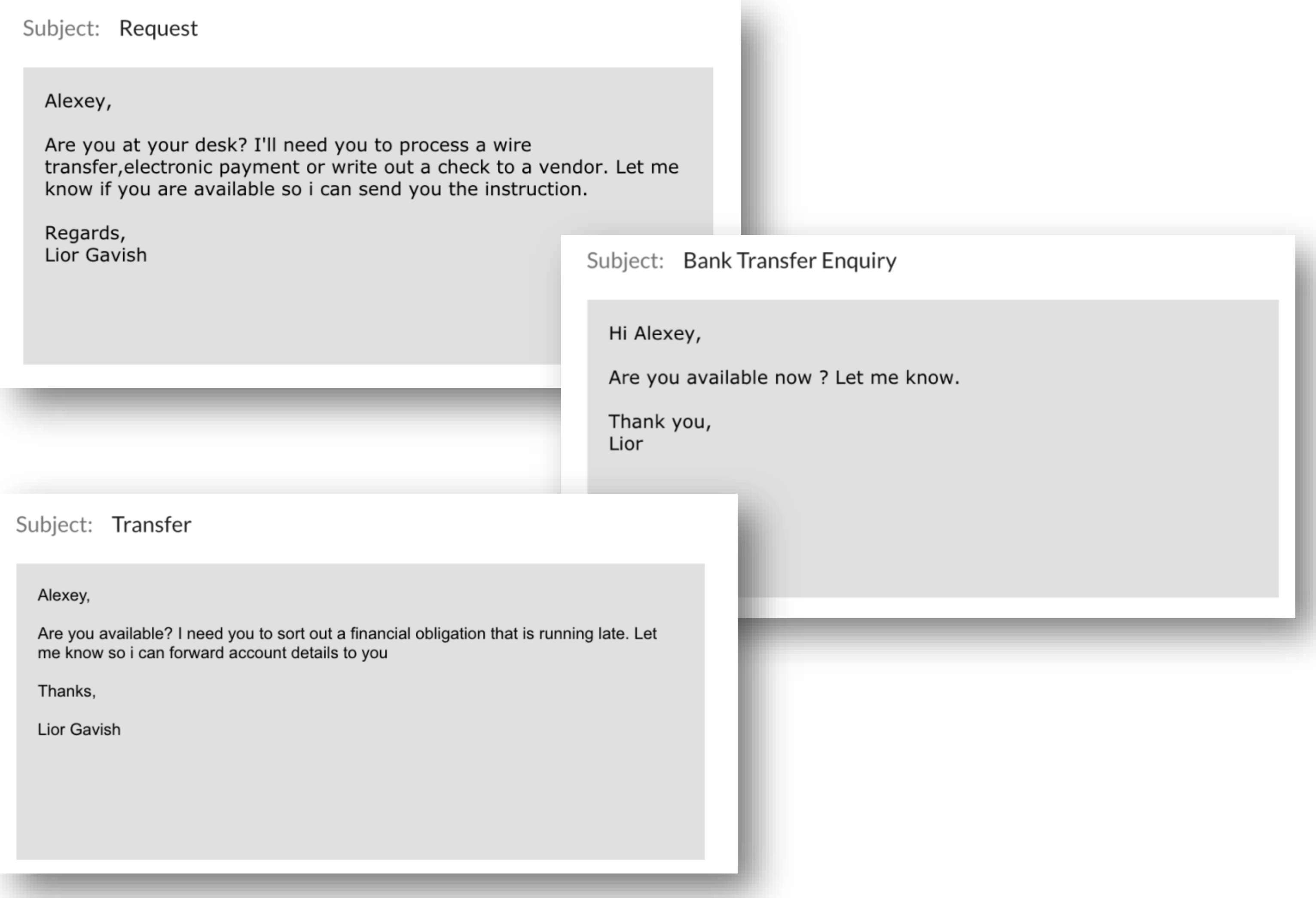
Question 1

(strategy, no code required)

In a wire fraud attack, an attacker will use impersonation to fool the target into wiring money to the attacker's account. To accomplish that, the attacker will send an email and pretend to be a colleague of the target, and ask for a payment to a vendor/consultant/etc. Please see attached examples of such attacks. Please note that the attacker will craft the email such that the FROM name is familiar to the target, but the FROM email address is actually controlled by the attacker. This allows the attacker to capture replies and correspond with the target without ever involving the impersonated sender.

For example (see screenshots), the attacker might send an email to Alexey and pretend to be his co-worker "Lior" by using the FROM email address “Lior Gavish <badguy@gmail.com>” instead of the actual work address (“Lior Gavish <lgavish@barracuda.com>”) or Lior's personal address (“Lior Gavish <lior.gavish@gmail.com>”).

Please suggest an approach that would allow us to automatically detect these attacks in real-time with good precision and recall. You can assume that you have the email corpus of 100 companies from the last 12 months, and that it contains ~1000 wire fraud attacks (though they are not labeled, so not readily available to you). Please describe not only your algorithmic approach, but also the process of developing your algorithm, evaluating it, and obtaining labeled data if you need it.



Subject: Request

Alexey,

Are you at your desk? I'll need you to process a wire transfer,electronic payment or write out a check to a vendor. Let me know if you are available so i can send you the instruction.

Regards,
Lior Gavish

Subject: Bank Transfer Enquiry

Hi Alexey,

Are you available now ? Let me know.

Thank you,
Lior

Subject: Transfer

Alexey,

Are you available? I need you to sort out a financial obligation that is running late. Let me know so i can forward account details to you

Thanks,

Lior Gavish

Question 2

(code question)

Please download the following dataset from http://ai.stanford.edu/~amaas/data/sentiment/ to get a list of 50,000 movie reviews:

- http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz

See folder "train/unsup" for the files. Write code to answer the following questions:

1. What are the 100 most commonly used words in the movie reviews?
2. Top 10 reviews closely related to review "train/neg/3316_2.txt"