ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ФРАНКА
Факультет електроніки та комп’ютерних технологій
Кафедра системного проектування

КУРСОВА РОБОТА

на тему:

“Виявлення і дослідження емоцій.”

Виконав:
студента групи ФЕІм-12с
Кравців Р. В
Науковий керівник:
Ляшкевич В. Я

Львів – 2023

Зміст

Зміст 2

Вступ 3

Розділ 1. Пов'язані роботи 4

1.1 Таксономії емоцій 4

1.2 Текстові корпуси для виявлення емоцій 5

1.3 Моделі виявлення емоцій NLP 7

Розділ 2. Виявлення емоцій у тексті 8

2.1 Визначення завдання 9

Розділ 3. Експериментальний дизайн 9

3.1 GoEmotions 10

3.2 Vent 11

3.3 Розробка багатоелементного бенчмарку 11

Розділ 4. Результати експерименту 14

4.1 Оцінка еталонних показників емоцій 14

4.2 Ієрархія емоцій у Vent 17

4.3 Експериментальна оцінка 19

Висновки 22

Посилання 23

Додаток А Деталі набору даних 25

A.1 GoEmotions 25

A.2 Vent 26

Додаток Б Дизайн еталонних показників 31

B.1 Підходи до представництва 31

B.2 Методи моделювання 32

Вступ

Ідентифікація емоційних сигналів є ключовою для вирішення низки завдань. Наприклад, виявлення емоцій необхідне для емпатичних чат-ботів, які можуть реагувати на емоційні потреби своїх користувачів. Розрізнення емоційно забарвленого контенту необхідне для вивчення вірусних, освітніх, політичних або підбурювальних взаємодій у соціальних мережах. Відстеження еволюції емоційного контенту, що надається користувачами, може допомогти запобігти домаганням або розробити ранні індикатори депресії. Міміка, мова, рухи тіла і текст є джерелами, з яких емоції можуть бути автоматично вилучені.

Аналіз емоцій відрізняється від аналізу настрою, який характеризує текст з точки зору полярності (позитивний, негативний або нейтральний). Для аналізу емоцій залучаються більший набір даних на які часто впливають такі аспекти, як амбіційність, непорозуміння, іронія чи сарказм.

Нещодавній прогрес у цій галузі став можливим завдяки успіху попередньо навчених мовних моделей, таких як Bidi- rectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), та випуску високоякісних великомасштабних анотованих наборів даних. У цій роботі я аналізую межі виявлення емоцій на основі тексту на прикладі двох найбільших джерел даних: GoEmotions (Demszky et al., 2020) та Vent (Lykousas et al., 2019). GoEmotions містить 58 тис. коментарів на Reddit, позначених, можливо, кількома тегами з 28 емоцій, анотованих читачами від третіх осіб. Відносно сирий корпус Vent (Lykousas et al., 2019 містить) містить 33 мільйони повідомлень, позначених однією з 705 емоцій первинними авторами від першої особи. Безпрецедентний обсяг цих наборів даних робить їх придатними для вивчення виявлення текстових емоцій у масштабі з різних точок зору.

Мета курсової роботи

Метою курсової роботи є дослідження та розробка методів та алгоритмів для автоматичного виявлення емоцій на основі аналізу тексту.

Для здійснення зазначеної мети служать наступні завдання:

- На базі двох найбільших та доступних на сьогоднішній день наборів даних GoEmotions та Vent навчити модель прогнозу емоцій;
- Провести тестування моделі на базі різних текстових повідомлень;
- Забезпечити зручний веб-інтерфейс для користувачів.

Розділ 1. Пов'язані роботи

Ми обговорюємо три сфери пов'язаної з цим роботи: існуючі таксономії для представлення емоцій, корпуси, що використовуються для побудови та оцінки систем виявлення емоцій, та підходи NLP, які можуть бути використані для реалізації систем виявлення емоцій на основі текстів.

1.1 Таксономії емоцій

Ландшафт людських емоцій був відображений кількома різними таксономіями та підходами. Екман (Ekman and Friesen, 1971) запропонував 6 основних емоцій, що виражаються через міміку в різних культурах: радість, смуток, гнів, здивування, відраза і страх. Незалежно від цього Плутчик (Plutchik, 1980) запровадив подібну таксономію, додавши до неї очікування та довіру, охарактеризувавши емоції за допомогою свого "Колеса емоцій" (Plutchik, 1980). Нещодавно були запропоновані більш деталізовані таксономії емоцій, які охоплюють високорозмірні зв'язки до 600 різних емоцій, кластеризуючи концепції емоцій за допомогою методів машинного навчання (Cowen et al., 2019). Ці таксономії показують багатогранну природу емоцій у різних культурах у вокалі (Cowen et al., 2018), музиці (Cowen et al., 2020) або міміці (Cowen and

Keltner, 2019). Окрім цих дискретних категоризацій, інші моделі, такі як афективно-циркумплексна модель емоцій (Русселл, 2003), розглядають емоції як пропорції у трьох вимірах, а не як дискретні категорії: Валентність (позитивна чи негативна), активація (активна чи пасивна) та домінування (домінантна чи підпорядкована). Нарешті, невелика кількість робіт розширює ці таксономії, щоб включити перспективу відправників і одержувачів емоційної комунікації (Ptaszyn- ski et al., 2009; Mohammad and Turney, 2013; Buechel and Hahn, 2017b,a).

У цій роботі ми зосереджуємося на категоріальних підходах і нещодавніх емоційних таксономіях, що охоплюють багатий спектр емоцій з точки зору відправників і одержувачів.

1.2 Текстові корпуси для виявлення емоцій

Для побудови систем виявлення емоцій фахівцям потрібні текстові дані, що містять емоційні ознаки. Ранні роботи, такі як SentiStrength (Thelwall та ін., 2010) та ANEW (Nielsen, 2011), використовували лексичні асоціації для аналізу емоцій, визначаючи, чи є текст позитивним, негативним або негативним, і в якій мірі. Лексичні підходи можна використовувати в системах, що базуються на правилах, де слова
ються з сигналом настрою або емоції. Складні моделі на основі правил, такі як VADER (Hutto and Gilbert, 2014), покладаються на анотовані людиною сентименти слів, а також сленг, модифікатори, наголос або пунктуацію. Інші підходи виходять за межі полярності: LIWC (Tauschik and Pennebaker, 2010) представили марковані словники, які зіставляють слова з їхньою емоційністю та ймовірністю полярності. Аналогічно, EmoLex (Mohammad та ін., 2013) створив на основі краудсорсингу лексикон асоціацій "слово-емоція", який класифікує слова з точки зору почуттів та емоцій відповідно до таксономії Плутчика.

Інший підхід полягає в тому, щоб розглядати виявлення емоцій як проблему керованого навчання, з корпусами, що включають емоційну

інформацію, які відрізняються за розміром, обсягом і підходом до маркування (Bostan and Klinger, 2018). Перші набори даних, такі як Affective Text (Strap- parava and Mihalcea, 2007), були невеликими: 1250 заголовків, маркованих за валентністю та емоціями Екмана. Щоб зменшити зусилля зі збору даних, у деяких роботах досліджувалися корпуси, видобуті з багатих на емоції середовищ, таких як соціальні мережі. Наприклад, набір даних про емоції Crowdflower (Crowd- flower, 2016) містить 40 тис. твітів, які розкривають емоції Екмана: ентузіазм, веселощі, ненависть, ней- траль, любов, нудьга, полегшення та порожнеча. Аналогічно, EmoNet (Abdul-Mageed and Ungar, 2017) застосували дистанційне спостереження, маркуючи твіти за допомогою хеш-тегів серед "кіл" у "Колесі Плутчика". Нарешті, паралельно з нашою роботою, Малко та ін. (2021) використовували дані Vent про обмежену кількість емоційних кат-егорій, що перетинаються з емоціями Екмана, і дійшли висновку, що самоконтрольовані мітки Vent вказують на емоційний зміст, виражений у тексті, підтримуючи більш детальний аналіз вираження емоцій.

Якщо розпізнавання емоцій розглядати як задачу навчання, то набори даних є компромісом між розміром і якістю. Анотовані людиною набори даних охоплюють тисячі зразків, часто орієнтовані на невелику кількість емоцій, як, наприклад, SocialNLP 2019 EmotionX chal- lenge або Crowdflower (Strapparava and Mihalcea, 2007; Crowdflower, 2016; Shmueli and Ku, 2019). Такі підходи, як EmoNet (Abdul-Mageed and Un- gar, 2017), можуть включати мільйони зразків, зібраних із соціальних мереж з використанням дистанційно контрольованих міток, що дозволяє створювати більші набори даних на момент публікації. Однак набори даних, зібрані з соціальних мереж, можуть бути приватними, оскільки прямий обмін даними часто заборонений, а контент регулярно видаляється, що обмежує відтворюваність. Доповнення
2561

Крім того, ярлики, створені шляхом дистанційного спостереження за допомогою хештегів, можуть не відповідати тому, як люди зазвичай сприймають або виражають емоції в різних сферах.

Наша робота спрямована на усунення цих недоліків шляхом аналізу двох найбільших наборів даних емоцій, анотованих людиною, GoEmotions і Vent, другий з яких містить 33 мільйони повідомлень, анотованих на 705 емоцій їхніми авторами в соціальній мережі. Ми описуємо ці два корпуси в розділах 4.1 та 4.2.

1.3 Моделі виявлення емоцій NLP

Підходи до виявлення емоцій часто ускладнюються труднощами вилучення емоційних сигналів з невеликих колекцій маркованих даних (Al-swaidan and Menai, 2020; Acheampong et al., 2020), використовуючи як інженерні, так і нейронні моделі. Деякі моделі на основі ознак вводять емоційні пріоритети, використовуючи асоціації "слово-емоція" як ознаки для контрольованих класифікаторів (Mo- hammad та ін., 2013). Інші підходи використовують статистичні методи, такі як Bag-of-Words, для представлення документів, покладаючи зусилля з вивчення емоційних асоціацій на модель (Silva et al., 2014).

Нещодавно більші набори даних дозволили навчити нейронні моделі, які перевершують свої традиційні аналоги. Наприклад, EmoNet зібрав 1,6 мільйона твітів і навчив на цих даних моделі на основі ШНМ (Abdul-Mageed and Ungar, 2017). Після виходу BERT (De- vlin et al., 2019) з'явилося багато нових робіт, присвячених точному налаштуванню моделей-трансформерів для навчання на основі дефіцитних даних про емоції. Наприклад, найкращі моделі на конкурсі SocialNLP 2019 Emo- tionX Challenge (Shmueli and Ku, 2019) перевершили найкращу попередню існуючу модель на 19% у мікро-F1. У цьому напрямку GoEmotions представили доопрацьований базовий класифікатор BERT з декількома мітками з 46% макро-F1 для 28 можливих міток (Demszky et al., 2020). Вибір моделювання часто залежав від наявності даних: підходи на основі правил (Hutto and Gilbert, 2014; Tauschik and Pennebaker, 2010) та лексичні підходи (Mo- hammad et al., 2013), коли навчальні дані були занадто рідкісними для рішень на основі ML, ненейронні методи в міру масштабування наборів

даних (Silva et al., 2014), а останнім часом - попередньо навчені нейронні моделі та трансформовані архітектури. Видатні результати тонко налаштованих трансформаторів стали рушійною силою більшості нещодавніх робіт, зокрема SocialNLP 2019 Emo- tionX challenge (Shmueli and Ku, 2019), GoEmo- tions (Demszky et al., 2020), а також виявлення емоцій
та бенчмарки для аналізу настроїв (Acheampong та ін., 2021).

На цій спільній основі ми досліджуємо прості базові лінії на основі Transformer, в яких ми використовуємо BERT як шар представлення та включаємо полегшені моделі поверх контекстуалізованих вбудовувань, перевершуючи попередні базові лінії BERT на мікро- F1 на 11,8%. Однак ми також вивчаємо нейронні та не-трансформаторні методи, які залишаються популярними в промисловості. Наш внесок охоплює різні підходи на великих наборах даних про емоції з великими просторами міток і 58 тис. / 9,75 млн текстів-зразків. Замість того, щоб вузько порівнювати варіації конкретних архітектур, наприклад, тонко налаштовувати трансформантні мовні моделі, ми працюємо з різноманітними відомими методами, щоб допомогти практикам вибрати підхід до моделювання з точки зору прогностичної ефективності та складності моделі.

Розділ 2. Виявлення емоцій у тексті

Фрагмент тексту може мати кілька асоційованих емоцій у випадках неоднозначності або при вираженні кількох почуттів одночасно, як показано в Таблиці 1. Таким чином, ми розглядаємо проблему як можливість застосування декількох правильних ярликів до певного фрагмента.

Text	Emotion
Wow. I just read the synopsis, and thats really what happens.	Surprise
What do you think? If you look at my question above? Last thing I should do is to say sorry?	Confusion
And then everyone clapped and cheered.	Joy, Admiration

Таблиця 1: Позначені пари текст-емоція з набору даних GoEmo- tions. У третьому прикладі показано приклад із кількома емоційними асоціаціями.

2.1 Визначення завдання

Сформулюємо задачу розпізнавання емоцій як задачу класифікації за багатьма мітками. За вхідним текстом $s \in U*$ та набором N емоцій, наша задача полягає у створенні (навчанні) функції $\varphi : U* 7\rightarrow PN$, яка відображає s у незалежні ймовірності для кожної емоції y1, y2, ..., yN . Розгляд виявлення емоцій як задачі з декількома мітками дозволяє нам застосувати ту саму архітектуру до багатокласових наборів даних. Це також дозволяє врахувати амплітуду навіть у наборах даних, де очікується лише один вихідний клас, наприклад, Vent.

Розділ 3. Експериментальний дизайн

Тут ми коротко описуємо набори даних GoEmotions та Vent, наш багатомітоковий еталонний дизайн класифікації, а також підходи до представлення та моделювання, які слугують будівельними блоками для наших детекторів емоцій. Ми включили детальну інформацію про експериментальний дизайн, гіперпараметри, додаткові результати та аналіз даних у Додаткові матеріали.

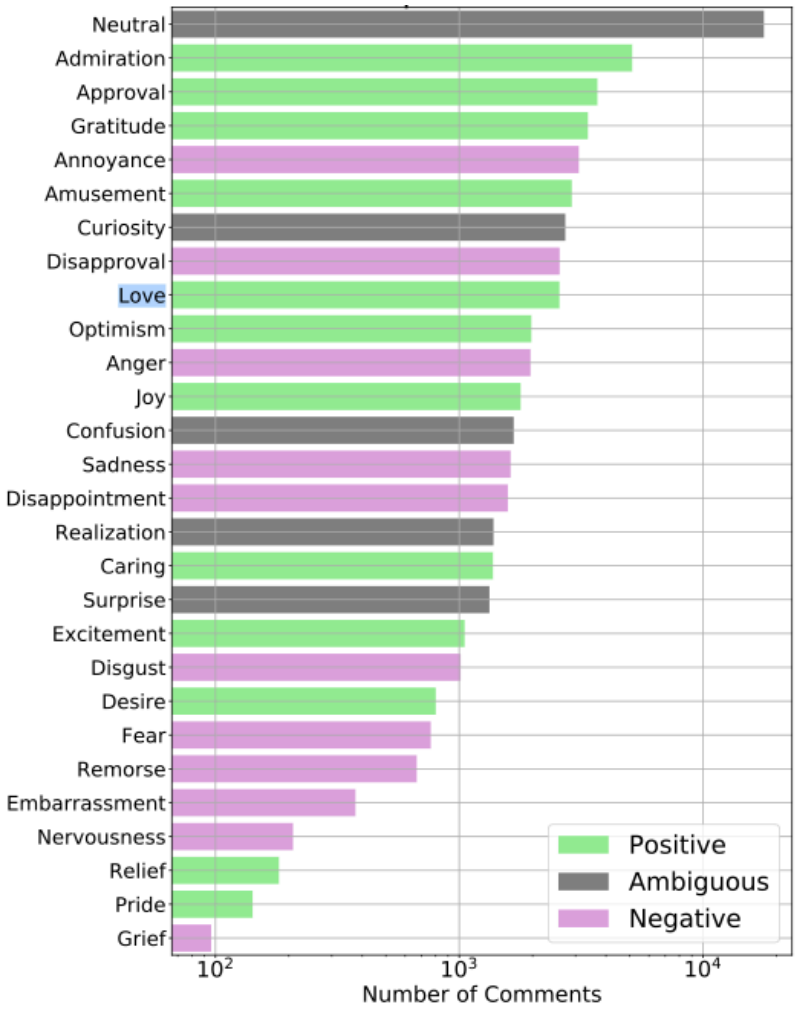


Рисунок 1: Частота міток GoEmotions та три категорії в (Demszky et al., 2020): Позитивні, Неоднозначні та Негативні.

3.1 GoEmotions

Набір даних GoEmotions (Demszky et al., 2020) містить 58 009 текстових фрагментів, зібраних з англомовних коментарів Reddit. Щонайменше 3 експерти віднесли кожен фрагмент до кількох емоцій з 27 категорій емоцій плюс нейтральна категорія, залишивши лише фрагменти, в яких 2+ експертів погодилися з одним або кількома ярликами. На рисунку 1 показано частоту та типи кожної емоції в наборі даних. Експертів просили позначати коментарі як "нейтральні", якщо вони не могли чітко визначити емоцію до коментаря. Набір даних супроводжується потужним базовим класифікатором BERT, побудованим

на основі BERT-бази (Devlin et al., 2019), який прогнозує 28 емоцій на основі контекстуального вбудовування останньої лексеми, досягаючи 0,46 макро-F1 і 0,51 мікро-F1 балів.

3.2 Vent

Набір даних Vent (Lykousas et al., 2019) містить понад 33 мільйони коментарів із соціальної мережі та мобільного додатку до неї, переважно англійською мовою. Кожен коментар або "вент" є самокоментарем автора, який ми будемо називати "вент", відповідно до його емоційного стану. Емоції структуровані на 63 категорії, що охоплюють 704 емоції.

На відміну від GoEmotions, де коментарі позначені емоціями "читача", які анотатори виводять з тексту, Vent - це набір даних емоцій "автора" - кожен коментар Vent позначений однією суб'єктивною емоційною міткою, наданою його автором. Набір даних надається як є, з мінімальною попередньою обробкою, анонімізуючи посилання на користувачів та URL-адреси. Для нормалізації корпусу ми (а) видалили стилістичні виділення, такі як курсив, (б) усунули зайві пробіли та (в) прив'язали посилання на користувачів та URL-адреси до спеціальних фіксованих токенів. Оскільки довжина Vents неоднорідна, ми обмежуємо довжину коментарів діапазоном від 3 до 32 лексем (75-й процентиль). Ми обмежуємося категоріями, які містять принаймні одну валідну емоцію, та ігноруємо емоції, позначені як відключені, які не використовуються принаймні раз на місяць або значення яких є неоднозначним, наприклад, ті, що містять емодзі на кшталт "Mushy". Відфільтровані дані містять 9,75 млн коментарів з 88 емоціями, які згруповані в 9 категорій.

3.3 Розробка багатоелементного бенчмарку

Ми розробляємо загальну архітектуру, яка дозволяє нам навчати, оцінювати і, за бажанням, передавати наші моделі, і використовуємо лише ті

методи, які можуть бути застосовані в потоковому міні-пакетному режимі. Наш підхід складається з трьох компонентів, показаних як послідовні кроки на Рисунку 2.

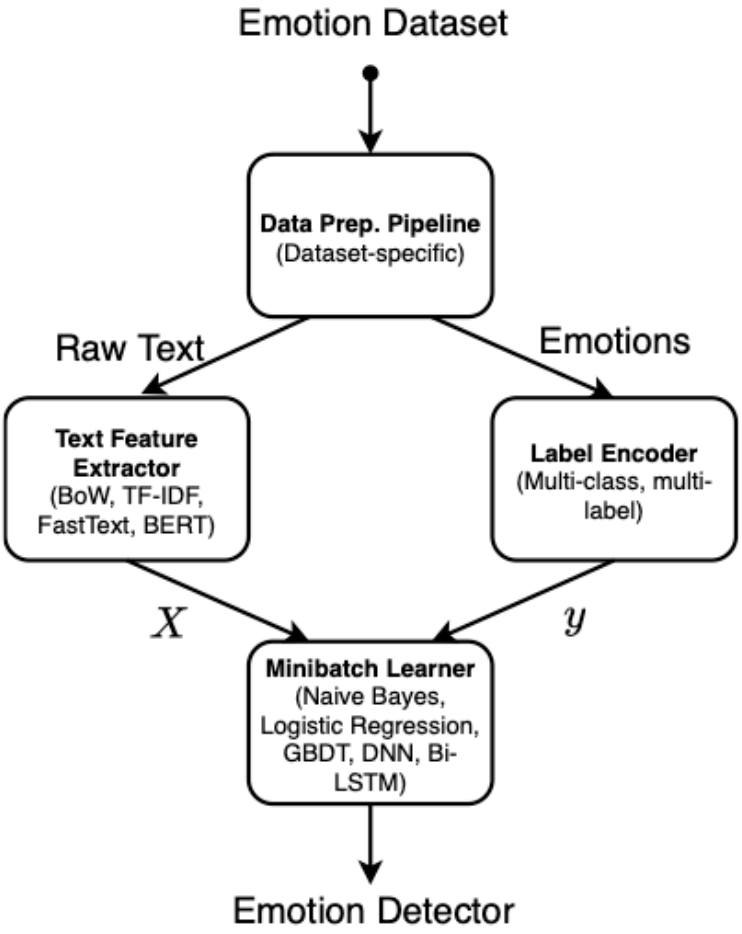
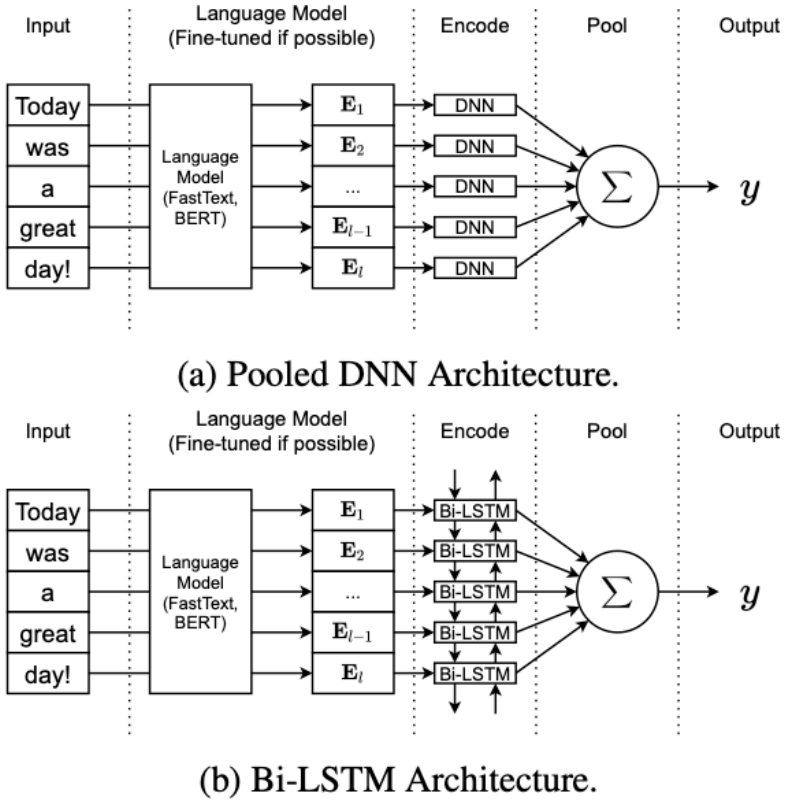


Рисунок 2: Високорівнева архітектура. Ми реалізували декілька модулів для прозорого виконання різних комбінацій представлення тексту та алгоритмів навчання. Варіанти в дужках детально описують альтернативні реалізації, доступні на кожному кроці тесту.

Ми реалізуємо два статистичні та два підходи до вбудовування представлення у поєднанні з п'ятьма різними алгоритмами навчання. Для статистичного представлення ми використовуємо моделі Bag-of-Words та TF-IDF, навчання наївного Байєса, логістичної регресії та збільшуваних випадкових лісів. Для

нейронних LM-представлень ми оцінюємо лише попередньо навчені англійські моделі, досліджуючи представлення на рівні слів за допомогою FastText та контекстуалізовані представлення за допомогою BERT з тонким налаштуванням.



Малюнок 3.Демонстрація нейронних архітектур у тесті. BERT точно налаштовано, а FastText – ні через обмеження реалізації

Як моделі для нейронних LM-зображень ми розробили дві прості нейронні архітектури, які отримують послідовності вбудовування на вході та об'єднують вихідні блоки для створення своїх виходів, як показано на Рисунку 3.

Ми реалізуємо задачу класифікації за кількома мітками, мінімізуючи середню втрату бінарної перехресної ентропії для N цільових емоцій.

1. (Pooled) Deep Neural Network. Ми застосовуємо ШНМ до кожного вбудованого токену незалежно та паралельно.

2. (Bi)-LSTM. Ми застосовуємо стекові (Bi)-LSTM, які споживають вбудовану послідовність послідовно один за одним.

Розділ 4. Результати експерименту

У цьому розділі ми вивчаємо результати нашого бенчмарку, оцінюємо поведінку навчених моделей та аналізуємо відмінності між емоціями автора та читача на Vent.

4.1 Оцінка еталонних показників емоцій

Спочатку ми оцінюємо методи виявлення емоцій, які ми представили в підрозділі 4.3. Ми зосередимося на чотирьох метриках класифікації з декількома мітками: Макро F1, мікро F1 та мікро усереднені точність і пригадування. Ми використовуємо оригінальні розбиття з наборів даних руху GoE, тоді як для навчання, валідації та тестування ми розбиваємо дані Vent на 80 / 10 / 10 розбиття, впорядковані за часом публікації. Наш аналіз описано окремо для кожного набору даних, а Таблиця 2 і Таблиця 3 містять результати для GoEmotions і Vent відповідно, причому моделі з найкращими показниками виділено жирним шрифтом.

Repr.	Model	M-F1	m-F1
BoW	N. Bayes	0.34	0.46
	Log. Reg.	0.45	0.53
	R. Forest	0.45	0.52
TF-IDF	N. Bayes	0.33	0.44
	Log. Reg.	0.47	0.53
	R. Forest	0.45	0.52
FT	DNN Pool	0.42	0.49
	Bi-LSTM	0.44	0.54
BERT	Baseline*	0.46	0.51
	DNN Pool	**0.48**	0.55
	Bi-LSTM	0.47	**0.57**

*Results as reported by (Demszky et a

Таблиця 2: Результати GoEmotions. Усі стандартні значення ≤ 0,01.

BERT-представлення перевершують всі інші конфігурації в обох наборах даних. Крім того, наші результати показують, що логістичні регресії перевершують наївні байєсівські та інкрементні випадкові ліси на конфігураціях з використанням Bag-of-Words або TF-IDF. На GoEmotions наш найвищий показник мікро-F1 становить 0,57, тоді як на Vent найвищий показник мікро-F1 становить 0,21.

Наші результати на GoEmotions перевершують попередні високі базові показники з використанням BERT, збільшуючи макро-F1 з 0,46 до 0,47 і мікро-F1 з 0,51 до 0,57. Позитивні результати свідчать про те, що наш підхід до розробки бенчмарку був доречним для досягнення надійного порівняння між нейронними та не-нейронними методами.

Repr.	Model	M-F1	m-F1	Pre	Rec
—	Random	0.02	0.04	0.02	0.69
BoW	N. Bayes	0.13	0.15	0.12	0.23
	Log. Reg.	0.13	0.15	0.12	0.20
	R. Forest	0.11	0.13	0.12	0.18
TF-IDF	N. Bayes	0.14	0.16	0.13	0.21
	Log. Reg.	0.14	0.16	0.14	0.21
	R. Forest	0.11	0.13	0.12	0.19
FT	DNN Pool	0.08	0.10	0.08	0.30
	Bi-LSTM	0.08	0.11	0.08	0.21
BERT	DNN Pool	0.17	0.19	0.16	0.24
	Bi-LSTM	**0.19**	**0.21**	**0.19**	**0.26**

Таблиця 3: Результати вентилювання. Всі stddevs ≤ 0.01, окрім моделей Fast- Text. Random - це базовий класифікатор, який виводить випадкову оцінку від 0 до 1 для кожного фрагмента та мітки.

Можливим поясненням підвищення продуктивності наших моделей порівняно з базовою лінією (Demszky et al., 2020) є те, що наш підхід об'єднує кожну лексему у вхідних даних, тоді як базова лінія використовує контекстне вбудовування останньої лексеми в речення для її передбачення. Ми спостерігаємо незначне покращення статистичних методів порівняно з базовою лінією, що може бути пов'язано з тим, що емоційна інформація здебільшого кодується на рівні слів.

На Vent ми бачимо, що модель BERT з об'єднаними Bi-LSTM шарами значно перевершує інші методи. У порівнянні з GoEmotions, не-BERT моделі працюють значно гірше, ніж моделі на основі BERT. Можливим поясненням цього є зашумлений характер вент-анотованого тексту з постів у соціальних мережах, вибір емоцій у яких іноді може бути довільним, а також підвищена завадливість 88 міток емоцій у Vent порівняно з 28 мішенями в GoEmotions.

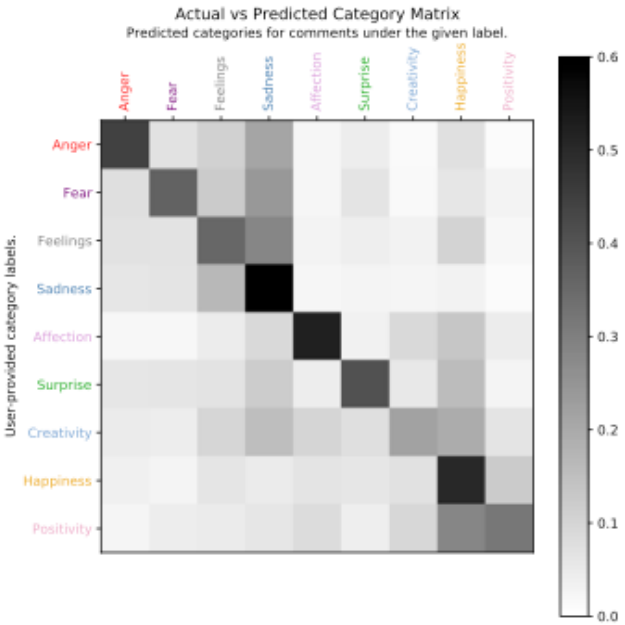


Рисунок 4: Нормалізована матриця плутанини між категоріями, що виконує прогнозування шляхом об'єднання емоцій як індикаторів категорії.

Коли ми варіюємо частку всієї навчальної вибірки, яка використовується для навчання, ми бачимо, що статистичні моделі працюють ближче до BERT, коли є менша кількість даних. Можливим поясненням такої поведінки може бути наш підхід до проектування, а саме, введення шарів поверх BERT, які сповільнюють збіжність, як показано на рисунку 3.

На наш подив, ми виявили, що FastText суттєво відстає від статистичних моделей. Зокрема, модель логістичної регресії TF-IDF значно перевершує FastText на Vent. Крім того, моделі TF-IDF, що використовують логістичні регресії або випадкові ліси показують подібні результати до моделі FastText з використанням об'єднаних Bi-LSTM на GoEmotions, незважаючи на відносну простоту моделей. Можливим поясненням спостережуваної різниці в продуктивності між Vent та GoEmotions може бути те, що попередньо навчені моделі FastText не здатні відтворювати сленг, друкарські помилки та специфічну для платформи лексику з Vent, незважаючи на те, що в основі їхньої моделі лежить n-грамова модель.

4.2 Ієрархія емоцій у Vent

Набір даних Vent містить одну емоцію, надану автором, на відміну від інших наборів даних, наприклад, GoEmotions, які надають анотації читачів. Однак наш підхід з використанням декількох міток дає ймовірності для кожної

емоції, що міститься в текстовому повідомленні. Це дозволяє нам аналізувати складну структуру емоцій, представлену нашою моделлю з Vent.

Замість того, щоб аналізувати парні кореляції між емоціями, як це зазвичай робиться, ми побудували (нормалізовану) матрицю плутанини M, де Mi,j - це частка спостережень прогнозованої емоції j, коли повідомлення про вентиляцію позначено i-ю емоцією. Отримана матриця плутанини на рівні категорій показана на рисунку 4 (матрицю плутанини на рівні емоцій див. у Додаткових матеріалах). Ми бачимо, що негативні категорії (Anger, Fear, Feelings, and Sadness) та позитивні категорії (Affection, Surprise, Creativity, Hap- piness, and Positivity) частіше "плутаються" всередині кожної групи, ніж між обома групами.

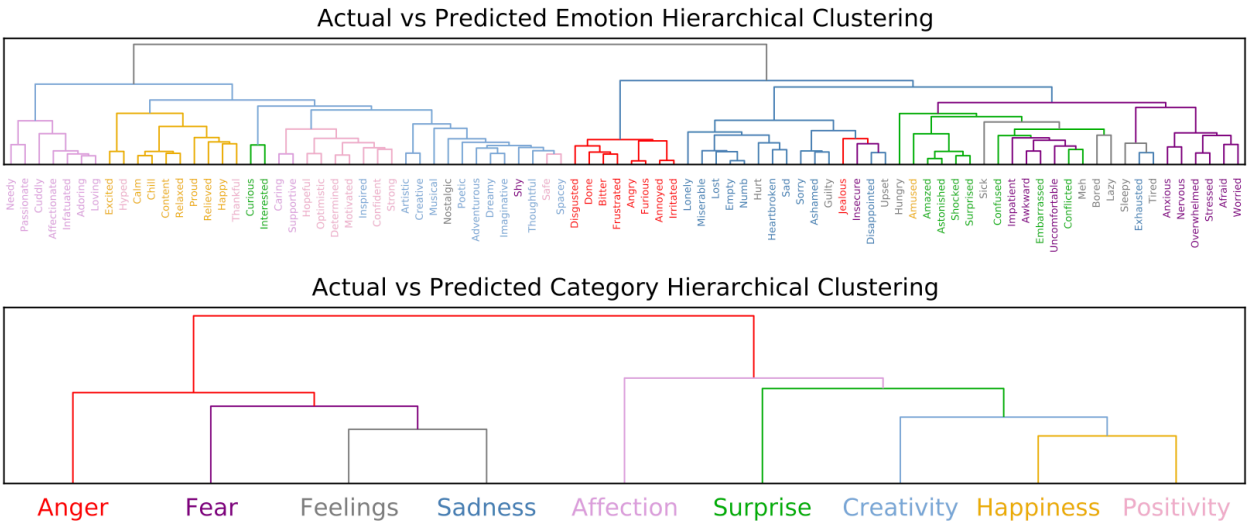


Рисунок 5: Дендрограми емоцій (вгорі) та категорій (внизу), отримані на основі нормалізованої активації для найкращої моделі, зафіксованої еквівалентом рівня емоцій на рисунку 4. Колір позначає категорію емоції.

Кожен рядок матриці M представляє патерн активації емоцій у цільовому просторі. Ми провели агломеративну кластеризацію, щоб виявити ієрархічну структуру такого простору. На рисунку 5 показано отриману дендрограму з використанням евклідової відстані між кожною парою рядків у M як метрики

зв'язку. Ми показуємо дендрограми на рівні емоцій та категорій, максимально об'єднуючи патерни активації між категоріями емоцій.
Ми бачимо, що кластери емоцій загалом збігаються з категоріями емоцій, незважаючи на те, що категорії не використовуються під час навчання. Верхні рівні ієрархії чітко розділяють позитивні та негативні емоції, тоді як наступні рівні забезпечують розумну ієрархію кластерів емоцій на різних шкалах. На рівні емоцій ми бачимо випадки, коли кластери містять емоції, що належать до різних категорій, але виявляються семантично значущими, наприклад, кластери, утворені Jealous, Insecure, Disappointed і Upset або кластер, утворений Sleepy, Exhausted і Tired на Рисунку 5.

4.3 Експериментальна оцінка

Ми розробили HIT-завдання за допомогою Amazon's Mechanic Turk (MTurk), щоб оцінити модель та відмінності між емоціями читача та автора.

4.3.1 HIT Design

Ми випадковим чином вибираємо 30 коментарів Vent з тестової вибірки для кожної з 88 емоцій, створюючи набір даних з 2640 коментарів. Ми нормалізуємо снайп-пети так само, як описано в розділі 4. Крім того, ми виключаємо коментарі, які містять терміни vent або nsfw (Not Safe For Work) щоб зменшити кількість самореферентного або недоречного контенту, з яким стикаються читачі. Коментарі групуються в тематичні групи (HIT) по 10 коментарів, за кожною з яких закріплюються 5 різних читачів, які проходять процедуру затвердження для забезпечення якості анотацій. Ми подали 264 HIT і отримали роботу від 84 різних читачів із середньою точністю емоцій 11,43% і середньою точністю категорій 34,26%.

На основі читацьких анотацій ми проаналізували узгодженість між анотаторами і виявили, що в середньому 1,81 (± 0,88) читачів з 5 читачів погоджуються з найчастіше присвоєною емоцією. На рівні категорій ми виявили, що в середньому 2,95 (± 1,03) читачів погоджуються з

найпоширенішим ярликом. Ми виявили, що більшість читачів можуть погодитися з категорією емоцій, але, як правило, не погоджуються з конкретними емоціями в межах категорії. Ми надаємо додаткову інформацію про узгодженість читачів між анотаторами в розділі E.4.2.

Модель перевершує людей-анотаторів (читачів) на підмножині фрагментів, поданих на розпізнавання на рівні категорій - 0,395 проти 0,383 мікро-Ф1, відповідно. Читачі демонструють вищий рівень пригадування (0,725 проти 0,472), але нижчу точність (0,261 проти 0,356). Це очікувано, оскільки ми не виконуємо додаткової пост-обробки, тобто фільтрації для вибору більшості міток, до читацьких анотацій.

Набір даних Vent зібрано з соціальної мережі з невідомими демографічними характеристиками користувачів, тому мітки емоцій, надані дописувачами, можуть бути зашумленими та/або двояко оціненими. Тому ми використовуємо читацькі анотації для дослідження придатності Vent для вивчення емоцій в масштабі.

Task	Label	Predictor	M-F1	m-F1
Emotion	Writer	Reader	0.151	0.151
		Model	0.181	0.181
	Reader	Model	0.241	0.246
		Writer	0.151	0.136
Category	Writer	Reader	0.383	0.383
		Model	0.395	0.395
	Reader	Model	0.467	0.471
		Writer	0.383	0.382

Таблиця 4: Порівняння результатів прогнозування міток читача або автора (N = 2640) за допомогою запропонованої нами моделі та іншої доступної перспективи.

Для цього ми інвертуємо завдання: ми передбачаємо анотації, надані читачами, за допомогою моделі, навченої на Vent. Ми намагаємося виміряти збіг між моделлю, читачами та авторами, оцінюючи, чи відповідає модель тому, що очікують читачі, більше, ніж оригінальні мітки Vent (тобто автори).

У Таблиці 4 показано оцінки макро- та мікро-F1 на рівні емоцій та категорій для моделі, читачів та авторів. Модель перевершує читачів у завданні передбачення письменників - тобто, коли ярлики дають письменники, а предикторами є або наша модель, або читачі - на 19,9% та 3,1% на емоційному та категорійному рівнях за відносним мікро-F1 (0,181 проти 0,151 та 0,395 проти 0,383 відповідно). Дивно, але модель також здатна передбачати неоднозначні емоційні ярлики читачів для тих самих текстів, незважаючи на те, що вона не була спеціально навчена цьому завданню. Зокрема, модель досягає на 35,9% вищого мікро-F1, коли оцінює читачів, а не авторів на рівні емоцій (0,246 проти 0,181), і на 19,2% на рівні категорій (0,471 проти 0,395). В якості контрольного експерименту ми також порівнюємо емоції письменників з емоціями працівників як предиктор емоцій працівників. Ми спостерігаємо мікро-F1 розрив, який може бути спричинений нерівномірними упередженнями читачів щодо певних емоцій, які впливають на підтримку кожної емоції1.

Наші результати показують, що моделі, навчені на великій кількості емоційних етикеток, наданих письменниками від Vent, здатні вловлювати емоції, які сприймають читачі. Наша модель досягає кращої продуктивності, коли вимірюється на читачах, а не на письменниках з точки зору мікро-F1, що узгоджується з існуючою літературою про виявлення емоцій на основі перспективи. Наприклад, у попередній роботі (Buechel and Hahn, 2017a) було виявлено, що перспективи читачів є кращими, ніж перспективи письменників, з точки зору узгодженості між анотаторами за моделлю емоцій VAD. Ці висновки можуть пояснити передбачувану природу анотацій працівників і нижчі

показники, які ми спостерігаємо при оцінюванні емоцій, наданих письменниками. Наші результати є першими, що враховують як авторську, так і читацьку точки зору на категоріальні таксономії у великих англійських корпусах, зібраних із соціальних мереж, що, наскільки нам відомо, не вивчалося в літературі з розпізнавання емоцій.

Висновки

Ми представляємо принциповий аналіз методів виявлення емоцій на двох найбільших доступних на сьогоднішній день наборах даних: GoEmotions та Vent. Наш ретельний бенчмарк показує, як поводяться різні моделі, причому архітектури на основі BERT незмінно досягають найкращої продуктивності на всіх наборах даних. Ми надаємо в оренду EmotionUI, веб-інтерфейс для дослідників, щоб вони могли вивчати наші моделі23, а також ділитися нашим кодом, інструментами, анотаціями та експериментальними даними4.

На GoEmotions наша найефективніша модель демонструє покращення порівняно з попередньою базовою лінією на 11,8% відносного мікро-F1. На Vent ми навчили модель, яка перевершує читачів у прогнозуванні емоцій, наданих авторами, на 19,9% відносного мікро-Ф1. Дивно, але ця ж модель показує кращі результати при оцінці емоційних ярликів, наданих читачами, на 35,9% відносно мікро-Ф1, незважаючи на те, що вона не була навчена цьому завданню. Наші результати показують, що моделі, навчені на Vent, перевершують читачів у завданні передбачення емоцій, наданих авторами, навіть якщо модель працює краще на емоціях, наданих читачами, а не на оригінальних істинних мітках, наданих користувачами Vent (авторами). Ці результати відкривають нові напрямки дослідження автоматичного виявлення емоцій між читачами та авторами в дискретних таксономіях емоцій, а також їх значення для систем виявлення емоцій. Наша робота припускає, що завдання передбачення емоцій, які автори прагнуть висловити, є складнішим, ніж виявлення емоцій, які сприймаються читачами. Однак більшість сучасних

досліджень з розпізнавання емоцій в анотованих текстах зосереджені на читацьких емоціях.

Посилання

Франциска Ачеампонг, Генрі Нуну-Менса та Венью Чен. 2021. Трансформаторні моделі для виявлення емоцій на основі тексту: огляд підходів, заснованих на бертах. Огляд штучного інтелекту, сторінки 1-41.

Франциска Адома Ачеампонг, Чен Венью та Генрі Нуну-Менса. 2020. Виявлення емоцій на основі тексту: Досягнення, виклики та можливості. Інженерні звіти, 2(7):e12189.

Гарет Джонс (Garethjns). IncrementalTrees: Он-лайн-дерева для навчання на основі Scikit-learn. https://github.com/garethjns/ IncrementalTrees.

Пьотр Бояновскі, Едуард Граве, Арманд Жулен і Томаш Міколов. 2017. Збагачення векторів слів інформацією про підслова. Transactions of the Association for Computational Linguistics, 5:135-146.

Лаура-Ана-Марія Бостан і Роман Клінгер. 2018. Аналіз анотованих корпусів для класифікації емоцій у тексті. У 27-й Міжнародній конференції з комп'ютерної лінгвістики.
Елої Брассар-Гурдо та Річард Хурі. 2019. Виявлення підривної токсичності за допомогою інформації про настрої. У Третьому семінарі з питань образливих висловлювань в Інтернеті.

Ніколас Карліні, Флоріан Трамер, Ерік Воллес, Метью Ягельський, Аріель Герберт-Восс, Кетрін Лі, Адам Робертс, Том Б. Браун, Дон Сонг, Уль- фар

Ерлінгссон, Аліна Опреа та Колін Раффель. 2020. Вилучення навчальних даних з великих мовних моделей. CoRR, abs/2012.07805.

Алан Коуен і Дахер Кельтнер. 2019. Що відображає обличчя: Картографування 28 емоцій, що передаються за допомогою природної міміки. Американський психолог, 75 років.

Ніколаос Лікусас, Константінос Пацакіс, Андреас Кальтенбруннер та Вісенс Гомес. 2019. Масштабний обмін емоціями: Набір даних про вентиляцію. 13-та Міжнародна конференція AAAI з блогів та соціальних медіа, 13(01):611-619.

Майк Телуолл, Кеван Баклі, Георгіос Палтоглу, Ді Кай та Арвід Каппас. 2010. Виявлення сили настрою в короткому неформальному тексті. Журнал Американського товариства інформаційних наук та технологій, 61:2544-2558.

Лаура-Ана-Марія Бостан і Роман Клінгер. 2018. Аналіз анотованих корпусів для класифікації емоцій у тексті. На 27-й Міжнародній конференції з комп'ютерної лінгвістики.

Елой Брассар-Гурдо та Річард Хурі. 2019. Виявлення підривної токсичності за допомогою інформації про настрої. У Третьому семінарі з ненормативної лексики в Інтернеті.

Свен Бухель та Удо Ган. 2017a. EmoBank: Вивчення впливу перспективи анотації та формату представлення на вимірний аналіз емоцій. На 15-й конференції Європейського відділення Асоціації комп'ютерної лінгвістики.

Свен Бухель та Удо Ган. 2017b. Читачі проти письменників, автори проти текстів: Як впоратися з різними перспективами розуміння тексту в емоційному анотуванні. У 11-му семінарі з лінгвістичного анотування, с. 1-12.

Додаток А Деталі набору даних

A.1 GoEmotions

Набір даних GoEmotions побудований на основі коментарів, вибраних зі звалища Reddit за період з 2005 по 2019 рік. Фрагменти фільтруються для того, щоб (а) зменшити кількість ненормативної лексики, образливого або дискримінаційного контенту, (б) зберегти постійну довжину коментаря менше 30 токенів із середнім значенням 12 токенів на коментар і (в) збалансувати настрої, емоції та популярність субреддітів. Власні імена та релігії маскуються за допомогою спеціальних символів, замінюючи їх у тексті на [NAME] і [RELIGION] відповідно.

# examples	58,009
# emotions	27 + neutral
# unique raters	82
# raters / example	3 or 5
Marked unclear or difficult to label	1.6%
Labels per example	1: 83% 2: 15% 3: 2% 4+: .2%
# examples w/ 2+ raters agreeing on at least 1 label	54,263 (94%)
# examples w/ 3+ raters agreeing on at least 1 label	17,763 (31%)

Таблиця 5: Загальна статистика набору даних та анотацій для набору даних GoEmotions. Взято з (Demszky et al., 2020).

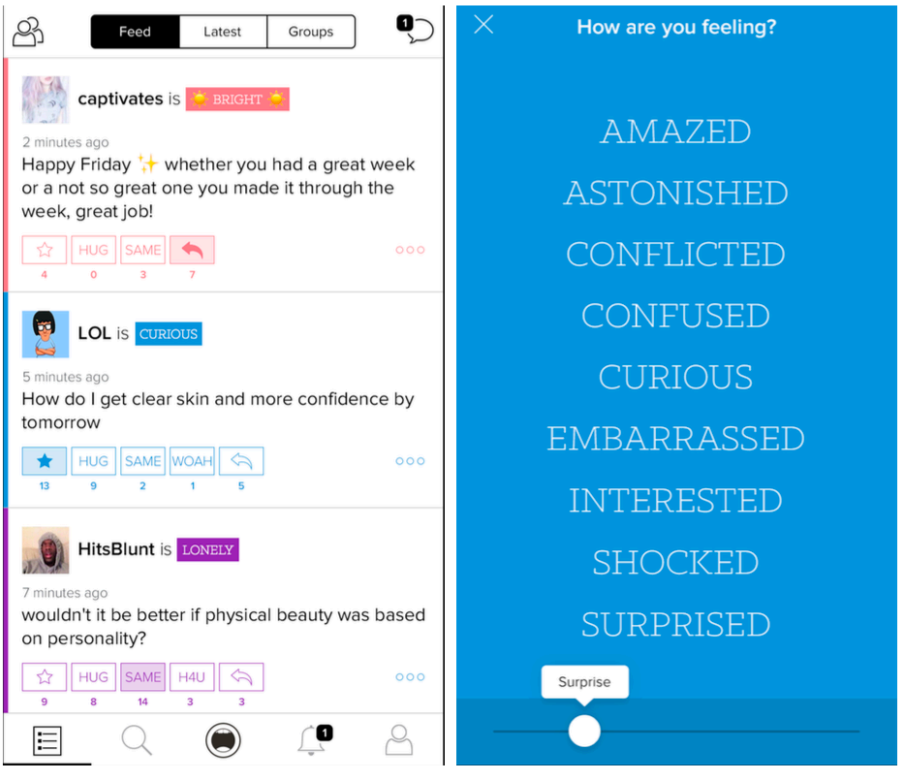
A.2 Vent

Набір даних Vent самокоментується користувачами (авторами) Vent за допомогою інтерфейсів, показаних на Рисунку 6. Розподіл довжини коментарів за кількістю токенів показано на Рисунку 7. Ми використали 32 як максимальну кількість токенів на один Vent відповідно до 75-го процентиля у розподілі довжини.

A.2.1 Процедура фільтрації

Ми відфільтрували набір даних, щоб (а) залишити місяці з достатнім обсягом і (б) залишити лише стабільні емоції, які з'являються щомісяця, починаючи з обраної точки.

На Рисунку 8 ми показуємо розподіл емоцій у часі з точки зору загальної частоти емоцій з кожної з 9 категорій емоцій. У роботі (Lykousas et al., 2019) автори повідомляють, що активність користувачів досягає піку в квітні 2015 року, а потім повільно знижується.



(a) Vent Feed (b) Emotion Picker

Рисунок 6: Скріншоти інтерфейсу програми Vent, взяті з (Lykousas et al., 2019).

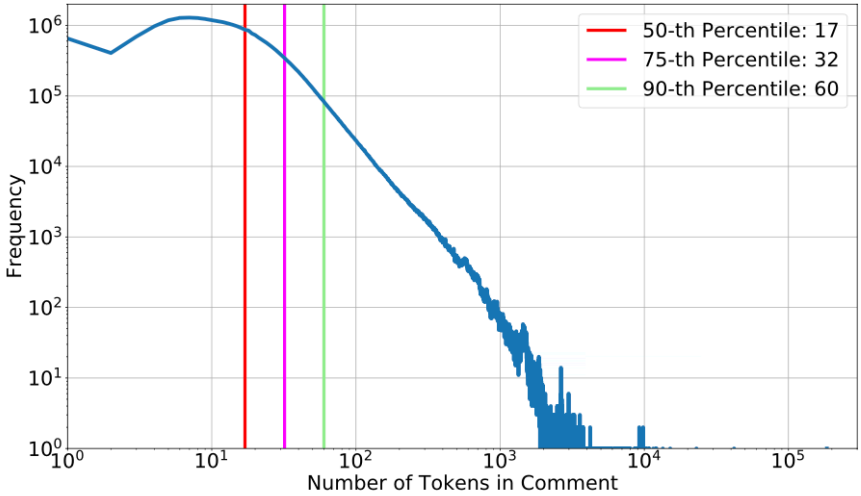


Рисунок 7: Розподіл довжини коментаря у кількості токенів.

Ми обираємо фільтр у липні 2016 року за двома критеріями: по-перше, ми обчислюємо кількість валідних виразних емоцій з плином часу (визначення валідної емоції див. в основному тексті). Це показано на рис. 9 (синя лінія). Ми бачимо, що після липня 2016 року було створено переважну більшість виразних емоцій. По-друге, щоб охарактеризувати їхню стабільність, ми обчислили відносну ентропію між розподілами емоцій двох послідовних місяців. Це показано на рисунку 9 (червоні лінії). Розподіли емоцій показують перетин (пунктирні червоні лінії) або об'єднання (суцільні червоні лінії) емоцій у кожній парі послідовних місяців. Ми бачимо, що після нашого відсікання в липні 2016 року обидві криві показують, що відносна ентропія приймає регулярні значення нижче 0,02, що вказує на те, що використання емоцій є стабільним після цього місяця.

A.2.2 Аналіз непристойних слів

На відміну від інших робіт, таких як GoEmotions (Demszky та ін., 2020), ми не виключаємо контент з нецензурною лексикою мову з навчального процесу. Наше обґрунтування полягає в тому, що Vent - це набір даних, зібраний з модерованої соціальної мережі. У цьому сенсі ми припускаємо, що модератори видаляють токсичну або неприйнятну поведінку, а також припускаємо, що нецензурна лексика є сильним емоційним сигналом, який необхідний для фіксації емоційних повідомлень у певних категоріях. Однак ми зазначаємо, що фільтруємо вміст nsfw (небезпечний для роботи) під час виконання завдання з анотування, щоб не наражати читачів на неприйнятний вміст.

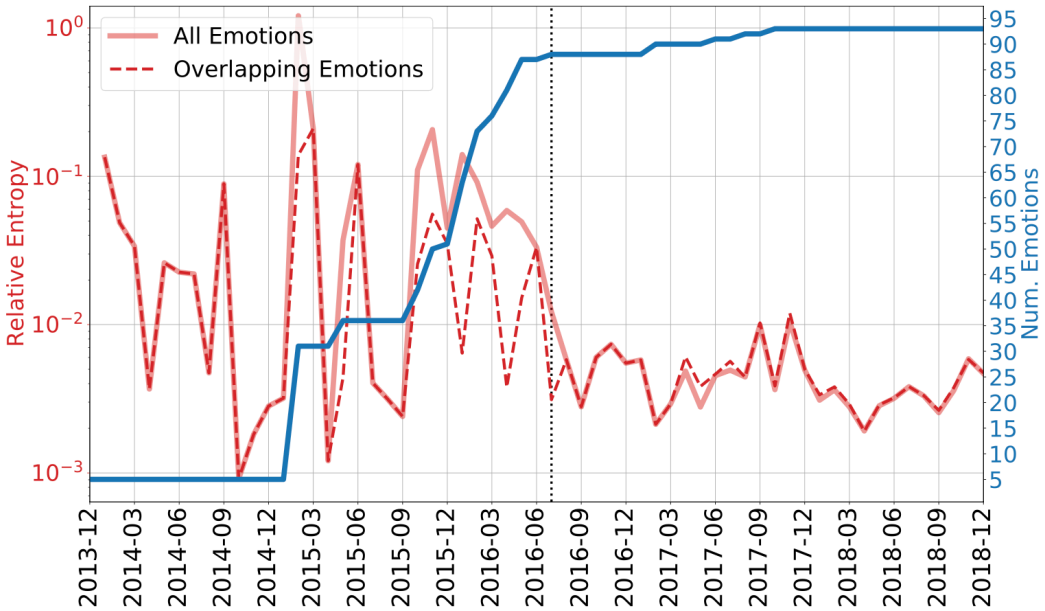


Рисунок 9: Відносна ентропія та кількість окремих емоцій місяць за місяцем. Відносна ентропія всіх емоцій включає піки, коли з'являються нові емоції, тоді як емоції, що перекриваються, відображають стабільність емоцій, які зберігаються між двома місяцями. Кількість емоцій зростає з часом, оскільки нові емоції вивільняються і замінюються новими. Ми бачимо, що кількість емоцій та відносні ентропії стабілізуються після 2016-07 рр., що показано пунктиром.

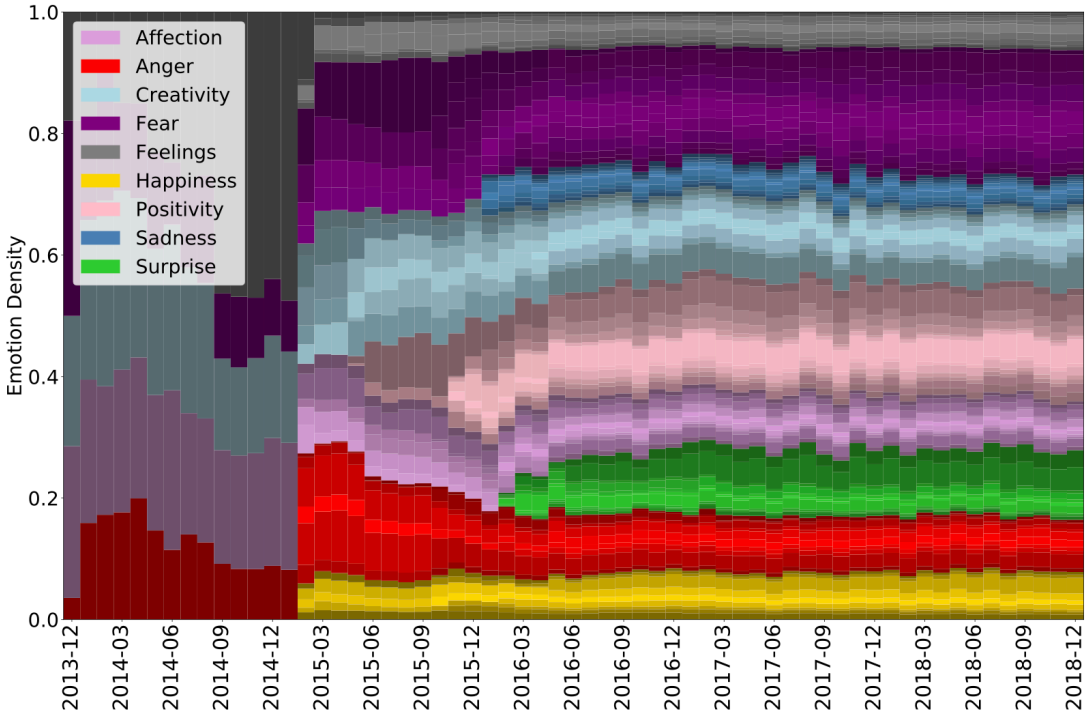


Рисунок 8: Щільність на рівні категорій щомісяця. Кожен колір представляє окрему емоцію, а його загальний розмір - це відсоток емоції в усіх повідомленнях цього місяця. Спочатку щільність є нерівномірною, доки кількість емоцій на плато додатку та всіх категорій не стане визначеною.

Щоб обґрунтувати рішення залишити непристойні терміни, ми досліджуємо, чи існують значні відмінності в ступені непристойності для кожної емоції та категорії. Ми визначаємо, чи є коментар непристойним, шляхом виявлення 763 слів з двох загальнодоступних списків непристойних слів 5 і лайливих слів 6. Ми виключили з нашого дослідження частотний аналіз: замість того, щоб враховувати частоту вживання нецензурних слів, ми використовуємо бінарний індикатор, який сигналізує про те, чи спостерігається непристойне слово з обох списків у коментарях Вента.

Ми спостерігаємо значні відмінності ($p < 10\text{-}20$, використовуючи бутстрап z-тести) у відсотках коментарів, що містять нецензурні слова, між категоріями, оцінені за допомогою бутстрапу на 100 незалежних прогонах з 10% набору

даних, відібраних за допомогою заміни. Ми виявили, що 10% коментарів у категорії "Позитив" містять нецензурні слова (найнижчий показник), тоді як 32% коментарів у категорії "Гнів" використовують нецензурну лексику (найвищий показник). На рисунку 10 показано кількість фрагментів для кожної емоції та категорії, а також частку нецензурних коментарів для кожної емоції.

Додаток Б Дизайн еталонних показників

B.1 Підходи до представництва

Ми реалізуємо статистичні та вбудовані репрезентації. Зокрема, ми використовуємо 4 різні алгоритми:

1. Мішок слів. Кожен документ подається у вигляді розрідженого вектора, що містить кількість слів у ньому. Ми обмежуємо розмір словника після фільтрації англійських стоп-слів за допомогою списку Scikit-learn (Pedregosa та ін., 2011), але не застосовуючи жодних перетворень на кшталт стеммінгу. Наш словник містить лише найчастотніші слова, які ми налаштовуємо як гіперпараметр, показаний у Таблиці 6
2. TF-IDF. Кожен документ представлено у вигляді розрідженого вектора, що містить нормалізовану частоту терміна, поділену на зворотну частоту документа. Ми застосовуємо ту саму політику побудови словника, що й у Bag-of-Words, включаючи фільтрацію стоп-слів та використання найчастотніших слів, налаштованих як гіперпараметр, як показано в Таблиці 6.
3. FastText. Ми використовуємо попередньо навчену модель FastText (Bojanowski et al., 2017) для вбудовування послідовності токенів у речення. Ми обмежуємо довжину речення як гіперпараметр.
4. BERT. Ми використовуємо попередньо навчену англійську модель BERT (Devlin et al., 2019) для токенізації та вбудовування послідовності токенів у речення. Наш дизайн дозволяє градієнтам поширюватися, коли BERT використовується як вхідні дані для наступної нейронної моделі. Ми обмежуємо загальну довжину

токенів і можливість поширення градієнтів (наприклад, заморожування BERT) гіперпараметрами, як показано в Таблиці 6.

B.2 Методи моделювання

Ми працюємо з п'ятьма різними студентами:

1. Наївний Байєс. Ми використовуємо реалізацію Scikit-learn (Pe-
(Pe- dregosa et al., 2011) реалізацію багатономінального наївного Байєса в режимі "один проти решти" для класифікації з декількома мітками.
2. Логістична регресія. Аналогічно, ми використовуємо логістичну регресію на основі SGD від Scikit-learn у режимі "один проти решти" для класифікації за кількома мітками.
3. Інкрементні випадкові ліси. Ми використовуємо розширення до класифікатора випадкових лісів Scikit-learn для інкрементальної побудови лісу в міні-пакетному режимі, використовуючи бібліотеку incremental-trees Python (, Garethjns). Кожна міні-партія вирощує фіксовану кількість дерев, до максимальної кількості, для всіх партій.
4. (Об'єднані) нейронні мережі. Ми реалізуємо просту архітектуру DNN, яка застосовується до кожного емітованого токену паралельно з N вихідними нейронами, стільки, скільки є класів. Кінцевий результат обчислюється шляхом застосування функції об'єднання станів, як показано на рисунку 3a.
5. (Bi)-LSTM. Ми реалізуємо стекову (Bi)-LSTM архітектуру, яка споживає вбудовану послідовність у режимі seq2seq, опціонально обробляючи її в одному напрямку. Рекурентна мережа фіксує зв'язки між лексемами і виробляє N вихідних нейронів на кожну лексему. Кінцевий результат обчислюється шляхом застосування функції об'єднання до кожної лексеми, як показано на рисунку 3b.

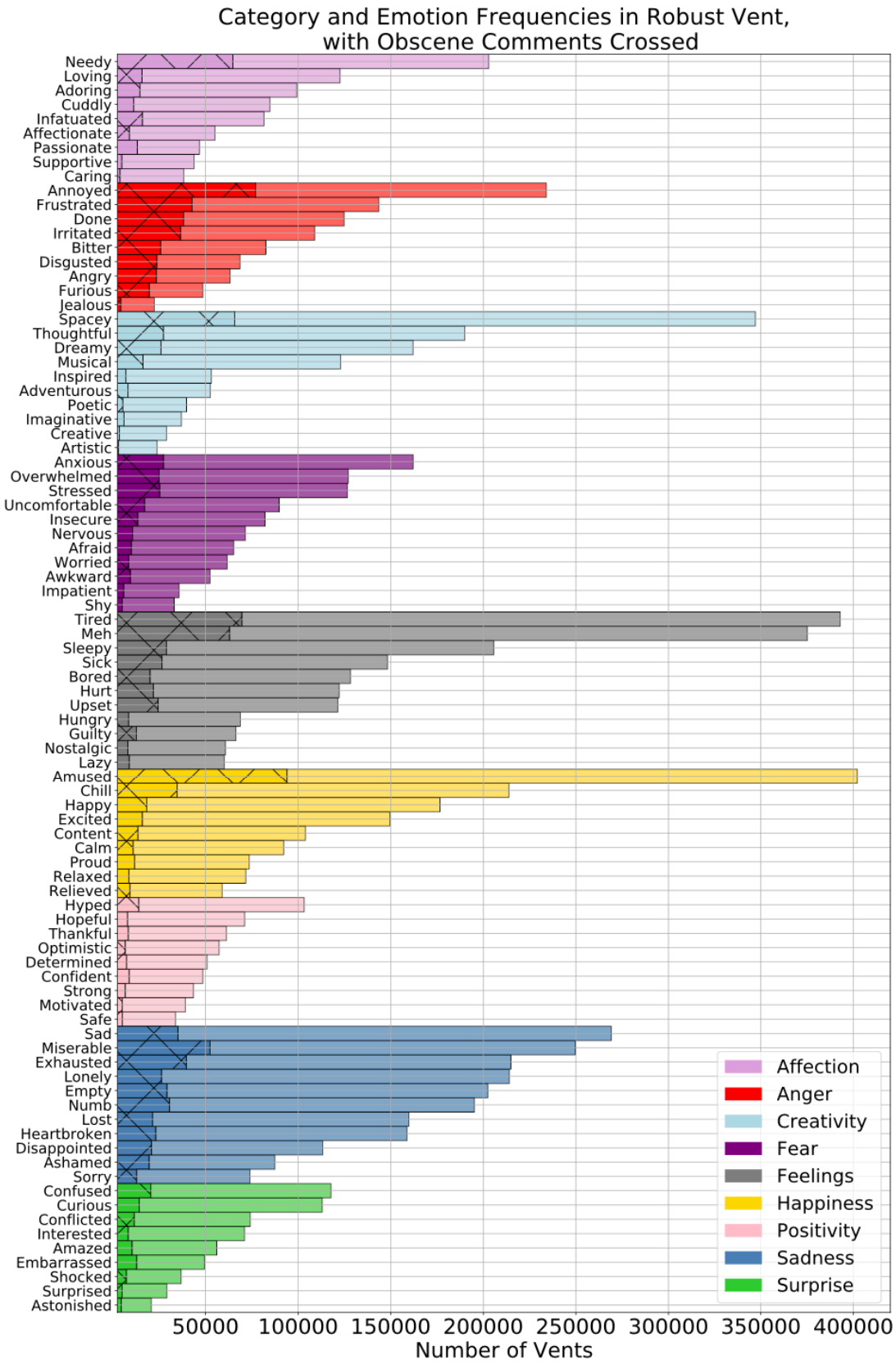


Рисунок 10: Емоційні позначки та їхня частота після фільтрації на відповідні емоції. Перехрещені темні області представляють частку даної категорії, яка містить принаймні один непристойний термін. Кольори показують категорії емоцій, які приблизно збігаються з категоріями Екмана: щастя, страх, смуток, здивування і гнів присутні, тоді як огида розглядається як емоція під гнівом. Є ще чотири додаткові категорії, які відповідають суб'єктивним станам вищого рівня: творчість, позитивність, почуття та прихильність.