Homework #5. Behaviour exploration improvement

Author: `Markiian Mandzak`

Total time spent on h/w (in minutes): ~360min

The original propaganda dataset by Kate Burovova should be already extracted into directory `./data/channels`

0. Setup

0.1. Imports

In [1]:
```
import json
import os
import re
import string
from collections import Counter

import altair as alt
import hdbscan
import langcodes
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from gensim.models import Word2Vec
from imblearn.under_sampling import RandomUnderSampler
from langid.langid import LanguageIdentifier, model as lang_model
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score
from sklearn.model_selection import train_test_split
import umap.umap_ as umap
import gender_guesser.detector as gender

alt.data_transformers.disable_max_rows()
```

```
/Users/markson/Desktop/UCU/UCU_6K1S_ComputationalSocialSciences/venv/lib/p
ython3.11/site-packages/tqdm/auto.py:21: TqdmWarning: IProgress not found.
Please update jupyter and ipywidgets. See https://ipywidgets.readthedocs.i
o/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
```

Out[1]:
```
DataTransformerRegistry.enable('default')
```

0.2. Constants

In [2]:
```
DIALOGS_MERGED_DATA_PATH = "../data_personal/merged_data/dialogs_data_all
DIALOGS_META_MERGED_DATA_PATH = "../data_personal/merged_data/dialogs_use
DICTS_DIR = "tone-dict-ukrainian/dicts"
MY_UID = 540076029
```

0.3. Stopwords

In [3]:
```
# https://github.com/skupriienko/Ukrainian-Stopwords/blob/master/stopword
uk_stopwords = {'а', 'аби', 'абиде', 'абиким', 'абикого', 'абиколи', 'аби
ru_stopwords = { "и", "но", "на", "что", "в", "он", "его", "её", "мы", "в
en_stopwords = {"'ll", "'tis", "'twas", "'ve", "10", "39", "a", "a's", "a
all_stopwords = uk_stopwords | ru_stopwords | en_stopwords
```

0.3. Sentiments

In [4]:
```
dict_lemmatized = {}
tone_dict = {}
for file_name in sorted(os.listdir(DICTS_DIR)):
    if file_name.startswith("dict_lemmatized"):
        print(f"Reading {file_name}")
        with open(os.path.join(DICTS_DIR, file_name), "r") as f:
            dict_lemmatized.update(json.load(f))
    elif file_name in ["tone-dict-ru.csv", "tone-dict-ua.csv"]:
        print(f"Reading {file_name}")
        td = pd.read_csv(os.path.join(DICTS_DIR, file_name), sep=";").set
        tone_dict.update(td)
```

```
Reading dict_lemmatized_en_words.json
Reading dict_lemmatized_ru_words.json
Reading dict_lemmatized_ua_words.json
Reading tone-dict-ru.csv
Reading tone-dict-ua.csv
```

In [5]:
```
emoji_sentiment_df = pd.read_csv("emoij_sentiment_data.csv")
emoji_sentiment_df['sentiment'] = (emoji_sentiment_df['Positive'] - emoji
emoji_tone_dict = emoji_sentiment_df.set_index('Emoji')['sentiment'].to_d
tone_dict.update(emoji_tone_dict)
```

In [6]:
```
tone_dict
```

Out[6]:

```
{'аббат': 0.3667,
 'аббревиатура': 0.0,
 'абзац': 0.0,
 'абонемент': 0.1757,
 'абонентный': 0.0,
 'абордажный': -0.1205,
 'абориген': 0.0,
 'аборт': -0.7832,
 'абракадабра': -0.0935,
 'абрикос': 0.584,
 'абрикосовый': 0.46,
 'абсолют': 0.4832,
 'абсолютизировать': 0.1136,
 'абсолютность': 0.4024,
 'абсолютный': 0.2297,
 'абстрактность': 0.0,
 'абстрактный': 0.0,
 'абстракция': 0.0,
 'абсурдность': -1.0,
 'абсурдный': -0.9,
 'абхазец': 0.0,
 'абхазский': 0.0,
 'авангард': 0.473,
 'аванс': 0.5308,
 'авантюра': -0.4909,
 'авантюризм': -0.2753,
 'авантюрист': -0.1538,
 'авантюристический': 0.1202,
 'авантюристка': -0.4315,
 'авантюрный': -0.0294,
 'аварийка': -0.0968,
 'аварийный': -1.0,
 'аварийщик': 0.0943,
 'авария': -1.0,
 'аватарка': 0.19,
 'август': 0.1864,
 'августовский': 0.0357,
 'авиабилет': 0.069,
 'авиагородок': 0.2358,
 'авиакасса': 0.1699,
 'авиакатастрофа': -0.9828,
 'авиакомпания': 0.3293,
 'авиаконструктор': 0.4179,
 'авиакрыло': 0.0,
 'авиалиния': 0.0,
 'авиапарк': 0.3333,
 'авиаперелёт': 0.1,
 'авиапочта': 0.1,
 'авиарейс': 0.0,
 'авиасалон': 0.2636,
 'авиатранспорт': 0.2984,
 'авиационный': 0.0,
 'авиация': 0.3378,
 'авитаминоз': -1.0,
 'авокадо': 0.4512,
 'австралиец': 0.0,
 'австралийский': 0.0,
 'австрийский': 0.0,
 'авто': 0.0806,
 'автобиография': 0.2742,
```

```
'автобус': 0.051,
'автобусный': 0.0,
'автовладелец': 0.4141,
'автовокзал': 0.0357,
'автогонка': 0.0357,
'автогонки': 0.3532,
'автограф': 0.4091,
'автодорога': 0.0,
'автодорожный': 0.0,
'автодром': 0.2909,
'автозаправка': 0.0,
'автозаправочный': 0.0,
'автокатастрофа': -1.0,
'автоклуб': 0.069,
'автокресло': 0.374,
'автоледи': 0.2235,
'автолюбитель': 0.1,
'автомагазин': 0.2317,
'автомагистраль': 0.0,
'автомагнитола': 0.1632,
'автомат': -0.1094,
'автоматизация': 0.5,
'автоматизировать': 0.413,
'автоматически': 0.0172,
'автоматический': 0.0,
'автоматный': 0.117,
'автоматчик': -0.1682,
'автомашина': 0.2846,
'автомобилестроение': 0.3605,
'автомобилестроительный': 0.0714,
'автомобилист': 0.0,
'автомобиль': 0.3203,
'автомобильный': 0.1,
'автомобильчик': 0.4919,
'автомойка': 0.1,
'автономизация': 0.31,
'автономный': 0.0,
'автопарк': 0.25,
'автопилот': 0.3254,
'автопортрет': 0.0,
'автопробег': 0.2742,
'автопроизводитель': 0.3333,
'автор': 0.3136,
'авторитет': 0.9,
'авторитетность': 0.9,
'авторитетный': 1.0,
'авторский': 0.3958,
'авторство': 0.5234,
'авторучка': 0.2304,
'авторынок': 0.0,
'автосалон': 0.1,
'автосервис': 0.25,
'автоспорт': 0.412,
'автостанция': 0.1699,
'автостоп': 0.0,
'автотранспорт': 0.4194,
'автотранспортный': 0.0,
'автоугонщик': -1.0,
'автоцентр': 0.0556,
'автошкола': 0.3849,
```

```
'агент': 0.0357,
'агентка': 0.0862,
'агентский': 0.0,
'агентство': 0.0833,
'агентша': 0.0422,
'агитировать': 0.0548,
'агония': -1.0,
'агрессивность': -1.0,
'агрессивный': -1.0,
'агрессия': -1.0,
'агроном': 0.2984,
'адаптационный': 0.431,
'адаптация': 0.5952,
'адаптировать': 0.9194,
'адвокат': 0.2815,
'адвокатский': 0.1,
'адвокатура': 0.2813,
'адекватность': 0.9706,
'адекватный': 1.0,
'аджика': 0.2341,
'адидас': 0.1216,
'админ': 0.199,
'административный': 0.0,
'администратор': 0.2909,
'администраторский': 0.0,
'администраторша': 0.0796,
'администрация': 0.1729,
'адмирал': 0.45,
'адреналин': 0.6282,
'адрес': 0.0,
'адресный': 0.0,
'адресовать': 0.0806,
'адресок': 0.0769,
'адски': -1.0,
'адский': -1.0,
'ажиотаж': 0.0769,
'азарт': -0.1448,
'азартный': -0.5893,
'азбука': 0.5373,
'азербайджанец': 0.0,
'азербайджанский': 0.0,
'азиат': 0.0,
'азиатский': 0.0,
'азот': 0.0,
'аист': 0.6,
'айкидо': 0.1,
'айсберг': 0.0,
'айтишник': 0.0,
'айфон': 0.3017,
'академгородок': 0.4208,
'академик': 0.5559,
'академический': 0.388,
'академия': 0.5579,
'акация': 0.4583,
'акваланг': 0.4023,
'аквалангист': 0.0806,
'аквамарин': 0.3,
'аквапарк': 0.7881,
'акварель': 0.4462,
'акварельный': 0.3974,
```

```
'аквариум': 0.0,
'аквариумный': 0.0172,
'аккаунт': 0.1,
'аккорд': 0.4328,
'аккордеон': 0.3496,
'аккумулирование': 0.4652,
'аккумулятор': 0.0833,
'аккуратненький': 1.0,
'аккуратненько': 0.9828,
'аккуратность': 1.0,
'аккуратный': 1.0,
'акробат': 0.434,
'акробатика': 0.3244,
'акробатически': 0.3889,
'акробатический': 0.316,
'акробатка': 0.4213,
'аксессуар': 0.3824,
'акт': 0.1346,
'актив': 0.8859999999999999,
'активация': 0.4375,
'активизироваться': 0.7051,
'активизм': 0.7653,
'активирование': 0.3952,
'активировать': 0.4556,
'активист': 1.0,
'активистка': 0.84,
'активность': 0.8158,
'активный': 1.0,
'актриса': 0.3678,
'актуализироваться': 0.5556,
'актуальность': 1.0,
'актуальный': 0.8,
'актёр': 0.4778,
'актёрский': 0.3136,
'актёрство': 0.4643,
'акула': -0.2534,
'акустический': 0.0,
'акушер': 0.4545,
'акушерка': 0.5179,
'акцент': 0.0625,
'акцентирование': 0.4029,
'акцентировать': 0.2311,
'акция': 0.6022,
'алгебра': 0.0702,
'алгебраический': 0.1,
'алгоритм': 0.0,
'аленький': 0.9179,
'алкашка': -1.0,
'алкоголизм': -0.9828,
'алкоголичка': -0.9146,
'алкогольный': -0.686,
'алконавт': -1.0,
'аллах': 0.0,
'аллегорически': 0.1404,
'аллергический': -0.7241,
'аллергия': -1.0,
'аллея': 0.4762,
'аллигатор': -0.2009,
'алмаз': 0.66,
'алмазик': 0.4744,
```

```
'алмазный': 0.4706,
'алогичность': -0.3462,
'алогичный': -0.3585,
'алоэ': 0.3136,
'алтайский': 0.0,
'алтын': 0.4184,
'алфавит': 0.5571,
'алфавитный': 0.1,
'алхимия': 0.2857,
'алый': 0.3015,
'алыча': 0.4459,
'альбинос': 0.201,
'альбомный': 0.0,
'альбомчик': 0.5656,
'альпинизм': 0.1,
'альпинист': 0.3644,
'альпинистка': 0.4435,
'альпинистский': 0.0,
'альтернатива': 0.5213,
'альтернативный': 0.3067,
'альтруизм': 0.7921,
'альфа': 0.3788,
'альянс': 0.4767,
'алюминиевый': 0.0,
'алюминий': 0.1509,
'амазонка': 0.3036,
'амбал': -0.6818,
'амбиция': 0.4727,
'американец': 0.0,
'американизация': 0.0,
'американка': 0.0,
'американский': 0.0,
'аморальность': -1.0,
'аморальный': -1.0,
'амплитуда': 0.0,
'амулет': 0.6081,
'анализ': 0.3952,
'анализирование': 0.5075,
'анализировать': 0.5859,
'аналог': 0.199,
'аналогичность': 0.2091,
'аналогичный': 0.1111,
'анальгин': 0.2576,
'ананас': 0.4857,
'ананасовый': 0.4231,
'анархистский': -0.6308,
'анархия': -0.7,
'анатомия': 0.1,
'ангар': 0.0,
'ангел': 1.0,
'ангелок': 1.0,
'ангелочек': 1.0,
'ангельски': 1.0,
'ангина': -0.9,
'английский': 0.0,
'англичанин': 0.0,
'англичанка': 0.0,
'андроид': 0.0172,
'анекдот': 1.0,
'аниматор': 0.5079,
```

```
'анимация': 0.34,
'аниме': 0.2419,
'анкета': 0.0,
'анкетирование': 0.0,
'аннулировать': -0.5,
'аномалия': -0.5,
'аномальность': -0.575,
'аномальный': -0.543,
'аноним': 0.0,
'анонимка': -0.2778,
'анонимность': 0.0,
'анонимный': 0.0,
'анорексия': -1.0,
'ансамбль': 0.4851,
'антагонист': -0.7247,
'антагонистический': -0.6515,
'антарктический': 0.0,
'антеннка': 0.2419,
'антибактериальный': 0.7091,
'антибиотик': 0.4512,
'антивирус': 1.0,
'антивоенный': 0.6667,
'антигерой': -0.8421,
'антидепрессант': 0.1255,
'антиквариат': 0.436,
'антикварный': 0.3473,
'антилопа': 0.371,
'антинаучный': -0.7018,
'антиправительственный': -0.4344,
'антирелигиозный': -0.2575,
'антисанитария': -1.0,
'антисептик': 0.6463,
'антоним': 0.0,
'анфас': 0.0,
'апартаменты': 0.4344,
'апельсин': 0.4286,
'апельсинный': 0.3361,
'апельсиновый': 0.5373,
'аплодировать': 1.0,
'аплодисменты': 1.0,
'апогей': 0.12,
'апокалипсис': -1.0,
'апорт': 0.2097,
'апостол': 0.6089,
'апостроф': 0.1,
'аппарат': 0.2368,
'аппаратура': 0.1,
'аппендикс': -0.3532,
'аппетит': 0.6304,
'аппетитность': 0.8365,
'аппетитный': 1.0,
'аппликатор': 0.3529,
'аппликация': 0.4174,
'апрель': 0.1923,
'апрельский': 0.2805,
'аптека': 0.4923,
'аптекарша': 0.4034,
'аптекарь': 0.4692,
'аптечка': 0.6395,
'аптечный': 0.0,
```

```
'араб': 0.0,
'арабика': 0.0556,
'арабский': 0.0,
'арахис': 0.4,
'арахисовый': 0.2105,
'арбалетчик': 0.0,
'арбуз': 0.5444,
'арбузный': 0.2436,
'аргумент': 0.5,
'аргументированно': 0.4535,
'аргументированность': 0.8276,
'аргументированный': 0.7734,
'аргументировать': 0.541,
'арена': 0.1122,
'аренда': 0.1964,
'арендатор': 0.1,
'арендный': 0.0,
'арендовать': 0.0172,
'арендодатель': 0.1,
'арест': -1.0,
'арестование': -0.8939,
'арестовать': -0.9,
'арестовывать': -0.9324,
'аристократ': 0.4851,
'аристократически': 0.4174,
'аристократический': 0.5373,
'аристократичность': 0.5315,
'аристократия': 0.229,
'аритмия': -0.9,
'арифметически': 0.0941,
'арктический': 0.0,
'армеец': 0.1,
'армейский': 0.0,
'армия': -0.1048,
'армрестлинг': 0.2159,
'армянин': 0.0,
'армянка': 0.0,
'армянский': 0.0,
'аромат': 0.6818,
'ароматерапия': 0.7353,
'ароматизация': 0.3793,
'ароматический': 0.5435,
'ароматичность': 0.9,
'ароматный': 0.9706,
'арочный': 0.0,
'арсенал': 0.0217,
'артиллерист': 0.0663,
'артиллерия': -0.1047,
'артист': 0.5859,
'артистически': 0.7632,
'артистический': 0.8056,
'артистичный': 0.9643,
'артистка': 0.1,
'артрит': -1.0,
'арфа': 0.3797,
'архаический': 0.025,
'архаичный': -0.0256,
'археолог': 0.3226,
'археологический': 0.0,
'археология': 0.3824,
```

```
'архив': 0.2573,
'архивирование': 0.2706,
'архитектор': 0.416,
'архитектурный': 0.2258,
'асимметрия': -0.0806,
'аскорбинка': 0.5714,
'аскорбиновый': 0.2222,
'аспирантка': 0.0455,
'аспирантура': 0.4252,
'аспирин': 0.3684,
'ассистент': 0.4153,
'ассистентка': 0.3049,
'ассистировать': 0.5,
'ассорти': 0.3571,
'ассортимент': 0.2043,
'ассоциация': 0.0,
'ассоциирование': 0.1,
'ассоциировать': 0.1964,
'ассоциироваться': 0.1757,
'астероид': -0.0915,
'астральный': 0.0847,
'астрологический': 0.0,
'астрология': 0.2692,
'астронавт': 0.3607,
'астроном': 0.5301,
'астрономический': 0.0,
'астрономия': 0.3702,
'астрофизический': 0.0769,
'асфальт': 0.0,
'асфальтовый': 0.1,
'асфальтоукладчик': 0.25,
'атака': -0.7308,
'атаковать': -0.6538,
'атеизм': -0.0205,
'атеистка': -0.2273,
'ателье': 0.339,
'атлантический': 0.0,
'атлас': 0.374,
'атлет': 0.5976,
'атлетика': 0.65,
'атмосфера': 0.0769,
'атомный': -0.1699,
'аттестат': 0.5153,
'аттестация': 0.5,
'аттракцион': 0.6707,
'ауди': 0.1,
'аудио': 0.0556,
'аудиозапись': 0.0,
'аудиокассета': 0.2317,
'аудиосистема': 0.0294,
'аудитория': 0.1466,
'аукать': 0.1452,
'аукцион': 0.0,
'аукционер': 0.2377,
'аукционист': 0.0,
'аукционный': 0.0,
'аул': 0.0,
'аутизм': -0.677,
'аутист': -0.475,
'афганский': 0.0,
```

```
'афера': -1.0,
'аферист': -1.0,
'аферистка': -1.0,
'афиша': 0.2623,
'афиширование': -0.1522,
'афишировать': 0.101,
'африканец': 0.0,
'африканский': 0.0,
'афроамериканец': 0.0,
'афроамериканка': 0.0,
'ахать': 0.0379,
'ахинея': -0.9,
'ахнуть': 0.0588,
'ацетон': -0.0894,
'аэродром': 0.0357,
'аэрозоль': 0.0,
'аэромобиль': 0.2925,
'аэроплан': 0.1606,
'аэропорт': 0.1,
'аэрофлот': 0.0676,
'баба': -0.55,
'баба-яга': -0.9242,
'бабай': -1.0,
'бабахнуть': -0.4047,
'бабища': -0.9,
'бабка': 0.2431,
'бабло': 0.0414,
'бабник': -1.0,
'бабочка': 0.684,
'бабулька': 0.5794,
'бабуля': 0.9828,
'бабуся': 0.9247,
'бабушка': 0.9247,
'бабушкин': 0.0357,
'бабёнка': -0.255,
'баг': -0.9,
'багаж': 0.0,
'багажник': 0.0172,
'багажный': 0.0,
'багроветь': -0.3317,
'багровый': 0.2353,
'бадминтон': 0.3095,
'база': 0.0769,
'базар': 0.0644,
'базарить': -0.8333,
'базовый': 0.219,
'базука': -0.4587,
'байдарка': 0.2864,
'байк': 0.0714,
'байкер': -0.0437,
'бакенбарды': 0.0,
'баклажан': 0.0,
'баклажанный': 0.0172,
'баклан': -0.3793,
'бактерия': -0.0558,
'бал': 0.4194,
'бал-маскарад': 0.9,
'балабол': -0.9,
'балаган': -0.9286,
'балагурство': -0.0379,
```

```
'баламутить': -0.9066,
'баланс': 0.0769,
'балансирование': 0.3244,
'балансировать': 0.4507,
'балансировка': 0.5075,
'балбес': -0.9643,
'балда': -0.995,
'балдёж': 0.9677,
'балерина': 0.5294,
'балеринка': 0.5,
'балет': 0.6429,
'балетный': 0.3978,
'балкон': 0.0106,
'балкончик': 0.375,
'балл': 0.4194,
'баллон': 0.0,
'баллончик': 0.0,
'баловать': 0.0945,
'баловаться': -0.2439,
'баловень': -0.9783,
'баловница': 0.0172,
'баловство': -0.6395,
'бамбук': 0.3293,
'бамбуковый': 0.2105,
'бампер': 0.0,
'банальность': -0.4344,
'банальный': -0.4211,
'банальщина': -0.7759,
'банан': 0.364,
'банановый': 0.3852,
'банда': -0.8514,
'бандана': 0.0,
'бандит': -1.0,
'бандитизм': -1.0,
'бандитка': -1.0,
'бандитский': -1.0,
'бандюган': -1.0,
'банить': -1.0,
'банк': 0.2009,
'банкет': 0.5698,
'банкетный': 0.4076,
'банкир': 0.1261,
'банкирша': 0.129,
'банковский': 0.0172,
'банковый': 0.0769,
'банкомат': 0.0,
'банкрот': -1.0,
'баннер': 0.0,
'баночка': 0.3108,
'бант': 0.0357,
'бантик': 0.5373,
'банька': 0.9,
'барабан': 0.244,
'барабанить': -0.0625,
'барабанный': 0.0,
'барабанчик': 0.1,
'барабанщик': 0.2895,
'барабанщица': 0.1,
'баран': -0.1321,
'бараний': 0.0,
```

```
'баранина': 0.3898,
'баранчик': 0.338,
'барахлишко': -0.2281,
'барахло': -0.7051,
'барахтанье': -0.1794,
'барахтаться': -0.2282,
'барашек': 0.6732,
'барбарис': 0.4449,
'барбекю': 0.516,
'барбос': 0.2559,
'бардак': -1.0,
'бариста': 0.3571,
'бармен': 0.2791,
'баронский': 0.05,
'баррикада': -0.0511,
'барсук': 0.3361,
'барсучий': 0.0,
'бартер': 0.0,
'бархат': 0.4268,
'бархатистость': 0.75,
'бархатистый': 0.7222,
'бархатный': 0.7258,
'барыня': 0.1,
'барышня': 0.5081,
'барьер': -0.125,
'барьерный': 0.0,
'бас': 0.1,
'бас-гитара': 0.4416,
'бас-гитарист': 0.4268,
'басистый': 0.0833,
'баскетбол': 0.5515,
'баскетболист': 0.0769,
'баскетболистка': 0.4194,
'баскетбольный': 0.0,
'баснописец': 0.3824,
'басня': 0.54,
'бассейн': 0.4375,
'бастовать': -0.9194,
'бастующий': -0.9194,
'басурманин': -0.7737,
'батальон': 0.1,
'батарейка': 0.0,
'батарейный': 0.0526,
'батарея': 0.0357,
'батон': 0.3906,
'батончик': 0.3571,
'батут': 0.4389,
'батюшка': 1.0,
'батя': 0.6609999999999999,
'бахила': 0.0,
'бахнуть': -0.5045,
'бахрома': 0.1346,
'бацилла': -1.0,
'бачок': 0.0,
'башенка': 0.4174,
'башка': -0.6714,
'башкир': 0.0,
'башмак': 0.1571,
'башмачник': 0.3276,
'башмачок': 0.371,
```

```
'башня': 0.0926,
'баюкать': 1.0,
'бдительность': 0.9,
'бдительный': 0.9,
'бег': 0.4139,
'беганье': 0.1938,
'бегать': 0.1,
'бегемот': 0.1944,
'беглец': -0.69,
'беглянка': -0.5174,
'беговой': 0.0172,
'беготня': -0.2521,
'бегство': -0.514,
'бегун': 0.3049,
'беда': -1.0,
'беднейший': -1.0,
'бедненький': -0.7368,
'беднеть': -1.0,
'бедность': -0.9375,
'беднота': -0.9,
'бедный': -0.9167,
'бедняга': -0.7571,
'бедняжка': -0.4596,
'бедняк': -0.8491,
'бедолага': -0.6429,
'бедро': 0.0,
'бедствие': -1.0,
'бежать': 0.2448,
'бежевый': 0.1792,
'беженец': -0.525,
'безаварийный': 0.9,
'безалаберность': -1.0,
'безалаберный': -1.0,
'безалкогольный': 0.4498,
'безальтернативность': -0.6391,
'безапелляционный': -0.3182,
'безбашенность': -0.5114,
'безбедный': 1.0,
'безбилетник': -1.0,
'безбилетница': -1.0,
'безбилетный': -0.7314,
'безбожие': -0.7212,
'безболезненность': 0.7566,
'безболезненный': 0.9112,
'безбородый': 0.0,
'безбрачный': -0.0806,
'безветренный': 0.1449,
'безвизовый': 0.1613,
'безвинность': 0.7162,
'безвкусие': -0.9,
'безвкусица': -0.9545,
'безвкусный': -0.5476,
'безвластие': -0.875,
'безвозвратный': -0.6667,
'безволие': -1.0,
'безволосый': -0.25,
'безвредность': 0.7742,
'безвредный': 0.9,
'безвременность': -0.2133,
'безвременный': 0.0428,
```

```
'безвыходность': -1.0,
'безвыходный': -1.0,
'безголовость': -1.0,
'безголовый': -1.0,
'безголосый': -0.7437,
'безграмотность': -1.0,
'безграмотный': -1.0,
'безграничность': 0.4213,
'безграничный': 0.0,
'безгрешный': 1.0,
'бездарность': -1.0,
'бездарный': -1.0,
'бездействие': -0.625,
'бездействовать': -0.6466,
'безделица': -0.6898,
'безделка': -0.5513,
'безделушка': 0.0,
'безделье': -0.7941,
'бездельник': -1.0,
'бездельница': -0.9054,
'бездельничать': -1.0,
'бездельный': -0.9,
'безденежный': -1.0,
'бездетность': -0.6795,
'бездетный': -0.4512,
'бездеятельный': -0.9,
'бездождье': -0.3831,
'бездоказательность': -0.3761,
'бездоказательный': -0.5,
'бездомный': -1.0,
'бездонный': 0.0,
'бездорожье': -0.75,
'бездумность': -1.0,
'бездумный': -0.8839,
'бездушие': -1.0,
'бездушный': -1.0,
'бездымный': 0.1804,
'безжалостность': -1.0,
'безжалостный': -1.0,
'безжизненный': -0.9255,
'беззаботность': 0.1073,
'беззаботный': 0.2135,
'беззаконие': -1.0,
'беззаконник': -0.9444,
'беззаконный': -1.0,
'беззастенчивость': -0.2031,
'беззастенчивый': -0.1884,
'беззащитность': -0.771,
'беззащитный': -0.3511,
'беззвучность': 0.0,
'беззвучный': 0.0,
'беззубый': -0.7377,
'безликий': -0.4211,
'безлистный': -0.199,
'безлунный': -0.0172,
'безлюдность': -0.375,
'безлюдный': -0.3036,
'безмозглость': -1.0,
'безмозглый': -1.0,
'безмятежность': 0.3697,
```

```
'безмятежный': 0.5161,
'безнадзорный': -0.9684,
'безнадёга': -1.0,
'безнадёжность': -1.0,
'безнадёжный': -1.0,
'безнаказанность': -1.0,
'безнаказанный': -0.9,
'безнал': 0.0,
'безналичный': -0.0806,
'безногий': -0.9079,
'безнравственность': -1.0,
'безобидность': 0.7564,
'безобидный': 0.6591,
'безоблачность': 0.6328,
'безоблачный': 0.5565,
'безобразие': -1.0,
'безобразник': -1.0,
'безобразница': -0.9098,
'безобразничать': -1.0,
'безобразность': -1.0,
'безопасность': 0.9643,
'безопасный': 1.0,
'безоружность': -0.0909,
'безоружный': 0.0161,
'безосновательный': -0.6667,
'безостановочный': 0.0,
'безответность': -0.9167,
'безответный': -0.6,
'безответственность': -1.0,
'безответственный': -1.0,
'безотказность': -0.0959,
'безотказный': 0.288,
'безотрадность': -1.0,
'безошибочность': 1.0,
'безошибочный': 1.0,
'безработица': -0.9146,
'безрадостный': -1.0,
'безразличие': -0.626,
'безразличность': -0.5735,
'безразличный': -0.4627,
'безразмерный': 0.0667,
'безрассудность': -1.0,
'безрассудный': -1.0,
'безрассудство': -1.0,
'безрасчётно': -0.3,
'безрезультатный': -0.9107,
'безрогий': -0.0806,
'безропотный': -0.1799,
'безрукавка': 0.0,
'безрукий': -1.0,
'безрукость': -0.9828,
'безударный': 0.0556,
'безудержный': 0.1682,
'безумец': -0.8727,
'безумие': -0.9,
'безумность': -0.9,
'безумный': -1.0,
'безумство': -0.9,
'безупречность': 1.0,
'безупречный': 1.0,
```

```
'безусловный': 0.2075,
'безуспешный': -1.0,
'безустанный': 0.5094,
'безусый': -0.0806,
'безучастие': -0.6074,
'безучастность': -0.5538,
'безучастный': -0.5,
'безызвестность': -0.459,
'безызвестный': -0.2653,
'безымянный': 0.0,
'безыскусность': -0.7232,
'безыскусный': -0.7513,
'безыскусственность': -0.2124,
'безыскусственный': -0.1849,
'безысходный': -1.0,
'бейдж': 0.0,
'бейджик': 0.1,
'бейсбол': 0.0806,
'бейсболист': 0.1606,
'бейсболка': 0.0,
'бейсбольный': 0.0,
'бекон': 0.4194,
'беленький': 0.0,
'белеть': 0.0,
'белиберда': -0.9,
'белизна': 0.3378,
'белковый': 0.0625,
'белобрысый': -0.1238,
'беловатый': 0.0172,
'беловолосый': 0.0,
'белогрудый': 0.2416,
'белозубый': 0.7778,
'белокаменный': 0.0769,
'белокочанный': 0.16,
'белокурый': 0.0,
'белолицый': 0.0,
'белорус': 0.0,
'белорусский': 0.0,
'белоснежность': 0.7119,
'белоснежный': 0.6527,
'белохвостый': 0.0,
'белочка': 0.6696,
'белый': 0.2727,
'бельевая': 0.0,
'бельишко': 0.0187,
'бельчонок': 1.0,
'бельё': 0.1,
'беляш': 0.9,
'бенгальский': 0.2368,
'бензин': 0.0,
'бензиновый': 0.0,
'бензобак': 0.0,
'бензовоз': 0.0,
'бензозаправка': 0.0,
'бензозаправщик': 0.0172,
'бензоколонка': 0.1346,
'бензопила': 0.0,
'бензохранилище': 0.0,
'бергамот': 0.3649,
'берег': 0.0,
```

```
'береговой': 0.0,
'бередить': -0.9,
'бережливость': 1.0,
'бережливый': 1.0,
'бережность': 1.0,
'бережный': 1.0,
'бережок': 0.4913,
'беременеть': 0.3967,
'беременная': 0.6972,
'беременность': 0.657,
'беременный': 0.6207,
'берет': 0.0,
'беретик': 0.4375,
'беречь': 1.0,
'беречься': 0.8832,
'берлинец': 0.0,
'берлинский': 0.0,
'берлога': 0.0,
'берёза': 0.5146,
'берёзка': 0.4449,
'берёзовый': 0.325,
'бес': -1.0,
'беседа': 0.4706,
'беседка': 0.3873,
'беседовать': 0.5806,
'бесить': -1.0,
'беситься': -1.0,
'бескомпромиссно': 0.0381,
'бескомпромиссность': -0.3717,
'бескомпромиссный': -0.1223,
'бесконечность': 0.0,
'бесконечный': 0.0,
'бесконтактный': -0.1875,
'бесконтрольность': -0.8590000000000001,
'бесконтрольный': -1.0,
'бескорыстие': 1.0,
'бескорыстность': 0.9444,
'бескорыстный': 1.0,
'бескрайний': 0.3797,
'бескрайность': 0.1,
'бескрайный': 0.0217,
'бескровный': -0.6143,
'бескультурье': -1.0,
'беспамятный': -0.8491,
'бесперебойность': 0.6316,
'бесперебойный': 0.4716,
'бесперспективный': -0.9107,
'беспечность': -0.6055,
'беспечный': -0.223,
'беспилотник': 0.1563,
'беспилотный': 0.0,
'беспламенный': -0.2562,
'бесплатность': 0.622,
'бесплатный': 0.8697,
'бесплодность': -0.9,
'бесплодный': -0.8143,
'бесповоротность': -0.6099,
'бесповоротный': -0.2714,
'бесподобность': 1.0,
'бесподобный': 1.0,
```

```
 'беспокоить': -0.9419,
 'беспокоиться': -0.4483,
 'беспокойный': -0.9107,
 'беспокойство': -0.8304,
 'бесполезность': -0.9519,
 'бесполезный': -1.0,
 'беспомощность': -0.9444,
 'беспомощный': -0.9,
 'беспорочно': 0.7718,
 'беспорядок': -0.8445,
 'беспорядочность': -1.0,
 'беспорядочный': -0.7368,
 'беспосадочный': -0.107,
 'беспочвенный': -0.425,
 'беспощадность': -1.0,
 'беспощадный': -1.0,
 'бесправие': -1.0,
 'беспредел': -1.0,
 'беспредельность': -0.6282,
 'беспредельный': -0.7091,
 'беспредельщик': -1.0,
 'беспрепятственный': 0.5579,
 'беспрерывный': 0.0085,
 'беспризорник': -0.9,
 'беспристрастие': 0.3793,
 'беспристрастность': 0.3429,
 'беспристрастный': 0.1914,
 'беспричинный': -0.1,
 'бесприютный': -1.0,
 'беспроглядный': -0.4815,
 'беспроигрышность': 1.0,
 'беспроигрышный': 0.9,
 'бессвязность': -0.6083,
 'бессвязный': -0.6696,
 'бессердечие': -1.0,
 'бессердечность': -1.0,
 'бессердечный': -1.0,
 'бессилие': -1.0,
 'бессильный': -0.9,
 'бесславность': -0.7035,
 ...}
```

In [7]:
```
dict_lemmatized
```

Out[7]:

```
{'расскажи': 'расскаж',
 'другое': 'другl',
 'зеркало': 'зеркало',
 'которе': 'котol',
 'https': 'https',
 'можно': 'можно',
 'так': 'old',
 'легко': 'легко',
 'переделать': 'переделать',
 'http': 'http',
 'надо': 'oldадо',
 'яндекс': 'яндекс',
 'там': 'old',
 'фсб': 'ФСфефс',
 'бекдор': 'бекдl',
 'sway': 'sway',
 'дефолтный': 'дефолтный',
 'самый': 'самый',
 'пиксельный': 'пиксельные',
 'никакие': 'никакой',
 'настройки': 'настройк',
 'работают': 'работают',
 'курсор': 'курсор',
 'мышной': 'мышнl',
 'походу': 'похiд',
 'никак': 'никак',
 'поменять': 'поменити',
 'хромиуме': 'хмиуее',
 'электроне': 'эктрое',
 'вяленде': 'венде',
 'установи': 'установа',
 'рефлектор': 'рлекте',
 'генерируй': 'герире',
 'зеркала': 'зеркало',
 'rsync': 'rsync',
 'найди': 'найди',
 'списке': 'списки',
 'удали': 'удати',
 'что': 'old',
 'стало': 'стати',
 'тебя': 'oldебя',
 'все': 'old',
 'пробовал': 'пробовал',
 'менять': 'менити',
 'работает': 'работает',
 'может': 'может',
 'быть': 'oldыть',
 'того': 'oldого',
 'аур': 'old',
 'тоже': 'oldоже',
 'жив': 'жити',
 'всё': 'old',
 'робит': 'робит',
 'норм': 'норм',
 'поменяй': 'поменять',
 'очевидно': 'очевидно',
 'только': 'тольl',
 'pacman': 'pacman',
 'или': 'old',
 'интернет': 'интернет',
```

```
'вообще': 'вообще',
'кстати': 'кстатин',
'сейчас': 'сейчl',
'репозитории': 'репозиторий',
'арча': 'oldрча',
'упали': 'упасть',
'пинг': 'пинга',
'доменному': 'денное',
'имени': 'имя',
'просто': 'просl',
'ping': 'ping',
'это': 'old',
'помогло': 'помогти',
'пробывал': 'пбывел',
'меня': 'oldеня',
'уау': 'уау',
'нём': 'він',
'ищет': 'ищет',
'тупо': 'тупо',
'результаты': 'результаты',
'реп': 'реп',
'частично': 'частично',
'aur': 'aur',
'лёг': 'льгагглькло',
'вроде': 'врогти',
'wat': 'wat',
'репе': 'oldепе',
'самое': 'самое',
'ждём': 'ждєнь',
'гонит': 'гонит',
'получается': 'получатися',
'тогда': 'тогда',
'conky': 'conky',
'manager': 'manager',
'rip': 'rippspsl',
'новым': 'новий',
'синтексисом': 'синтексисом',
'есть': 'oldсть',
'manager2': 'manager2',
'git': 'git',
'оказывается': 'оказыватися',
'исходники': 'исходник',
'кеше': 'кеш',
'версии': 'верси',
'раньше': 'раньl',
'были': 'oldыли',
'предупреждения': 'предупреждение',
'совсем': 'совсl',
'убрали': 'убрати',
'совместимость': 'совместимость',
'странно': 'странно',
'вот': 'old',
'недавно': 'навно',
'обновился': 'обновиться',
'отвалились': 'отвалилисе',
'темы': 'темы',
'говорит': 'горит',
'синтексис': 'стексе',
'тот': 'old',
'попробую': 'попробувати',
```

```
'google': 'google',
'cloudflare': 'cloudflare',
'лучше': 'лучше',
'кто': 'old',
'теме': 'ти',
'поймёт': 'понять',
'шутку': 'шутка',
'etc': 'etc',
'resolv': 'resolv',
'conf': 'conf',
'здесь': 'здесь',
'ошибка': 'ошибка',
'разрешения': 'разрешения',
'имён': 'им’ннннюн',
'попробовать': 'попробовати',
'изменить': 'изменити',
'dns': 'dns',
'сервер': 'сервер',
'лет': 'old',
'назад': 'назад',
'соединение': 'соединение',
'версия': 'версия',
'типа': 'тип',
'переход': 'переход',
'синтаксис': 'синтаксис',
'lua': 'lua',
'если': 'oldсли',
'нужно': 'нужно',
'mirrorlist': 'mirrorlist',
'reflector': 'reflector',
'тут': 'old',
'обновил': 'обновилолил',
'поднялся': 'подняться',
'кучу': 'кувати',
'как': 'old',
'используется': 'использоваться',
'граждане': 'гражданин',
'исправить': 'исправить',
'чую': 'чути',
'коньками': 'кькаеи',
'его': 'old',
'синтаксисом': 'синтаксис',
'беде': 'беда',
'давно': 'давно',
'сайт': 'сайт',
'такой': 'такой',
'тормознутый': 'тормознутый',
'стал': 'oldтал',
'bugzilla.kernel.org': 'bugzilla.kernel.org',
'последнее': 'последне',
'сообщение': 'сообщений',
'логе': 'oldоге',
'когда': 'когда',
'зависает': 'зисает',
'система': 'система',
'x86': 'x86',
'pat': 'pat',
'configuration': 'configuration',
'total': 'total',
'ram': 'ramum',
```

```
'covered': 'cover',
'16368m': '16368',
'found': 'found',
'optimal': 'ootimal',
'setting': 'setting',
'mtrr': 'mtrr',
'clean': 'clean',
'gran': 'gran',
'size': 'size',
'64k': '64',
'chunk': 'chunky',
'32m': '32',
'num': 'num',
'reg': 'reg',
'lose': 'lose',
'cover': 'cover',
'e820': 'e820',
'update': 'update',
'mem': 'mem',
'0xe0000000': '0xe0000000',
'0xffffffff': '0',
'usable': 'usable',
'reserved': 'reserved',
'поэтому': 'поэтий',
'нахрена': 'нахрен',
'вам': 'old',
'виртуалбокс': 'виртуалбокс',
'линуска': 'луска',
'квм': 'old',
'никакой': 'никакой',
'бокс': 'бокс',
'еще': 'old',
'памяти': 'памяти',
'хватает': 'хватает',
'точно': 'точно',
'указать': 'указати',
'прошлую': 'прошлую',
'загрузку': 'загрузка',
'системы': 'системы',
'похоже': 'похожеомодіб',
'делаю': 'делай',
'воспроизвёл': 'воспроизвёл',
'ошибку': 'ошибка',
'зависла': 'зависнути',
'намертво': 'нертео',
'перезагрузил': 'перезагрузил',
'харду': 'харду',
'ввёл': 'ввести',
'команду': 'команда',
'journalctl': 'journalctl',
'since': 'since',
'min': 'min',
'ago': 'ago',
'сообщения': 'сообщений',
'ошибке': 'ошибка',
'нет': 'old',
'запрашиваю': 'запрашивае',
'лог': 'лог',
'капец': 'капец',
'умный': 'умный',
```

```
'используй': 'использувати',
'браузер': 'браузер',
'для': 'old',
'редактирования': 'редактирование',
'текста': 'тексто',
'сказать': 'сзать',
'rsc': 'rsc',
'pycharm': 'pycharm',
'electron': 'electron',
'vim': 'vim',
'будет': 'будет',
'очень': 'очень',
'красив': 'краситисти',
'kde': 'ддededeedet',
'konsole': 'konsole',
'поставить': 'поставити',
'прозрачный': 'прозрачный',
'фон': 'фон',
'размытием': 'рмытие',
'попробуй': 'попробувати',
'ядро': 'ядро',
'память': 'память',
'кончается': 'кончаться',
'логах': 'лога',
'dmenu': 'dmenu',
'причем': 'причемрчочкуч',
'сделал': 'сделал',
'local': 'local',
'share': 'share',
'applications': 'applications',
'свой': 'свой',
'desktop': 'deskto',
'файл': 'файл',
'запуска': 'запуск',
'одной': 'одной',
'фигни': 'фигня',
'аргументами': 'аргумент',
'оно': 'old',
'отображается': 'отображаться',
'rofi': 'rofi',
'делать': 'делать',
'привет': 'привет',
'подскажите': 'подскажити',
'пожалуйста': 'пожалуйсте',
'чём': 'чём',
'проблема': 'проблема',
'при': 'old',
'запуске': 'запуск',
'виртуальной': 'виртуальный',
'машины': 'машина',
'virtualbox': 'virtualbht',
'гугл': 'гугл',
'несовместимости': 'несовместимости',
'ядром': 'ядр',
'линукс': 'линукс',
'стоит': 'стоит',
'дефолтное': 'долтне',
'переустанавливать': 'переустанавливать',
'размер': 'размер',
'шрифтов': 'шрифт',
```

```
'повлияет': 'повлиять',
'масштаб': 'масштабб',
'интерфейса': 'интерфейс',
'была': 'oldыла',
'нова': 'новий',
'был': 'old',
'нормальный': 'нормальный',
'поставил': 'поставил',
'нвидию': 'нвидl',
'уехал': 'уехать',
'нашёл': 'нашнтитинти',
'задаётся': 'задаваться',
'настройках': 'настройка',
'ide': 'ide',
'после': 'после',
'установки': 'установка',
'nvidia': 'nvidia',
'слишком': 'сшком',
'мелким': 'мелкий',
'таким': 'такий',
'сталкивался': 'сталкиваться',
'moshe': 'moshe',
'shelomov': 'shelomov',
'было': 'oldыло',
'перенесено': 'перенести',
'archlinuxofftopicru': 'archlinuxofftopicru',
'пользователь': 'пользователь',
'генту': 'гент',
'оскорблен': 'оорбле',
'тем': 'old',
'арчеры': 'арчel',
'считаются': 'считавтися',
'более': 'более',
'самодовольными': 'самодовольными',
'чем': 'old',
'кого': 'oldого',
'послал': 'послалоглумтиг',
'опять': 'опять',
'пропустил': 'пропустил',
'забанил': 'занил',
'правда': 'правда',
'могу': 'мога',
'даже': 'oldаже',
'глянуть': 'глянути',
'хех': 'хехк',
'третий': 'третl',
'блять': 'бити',
'раз': 'old',
'неделе': 'недтии',
'мразь': 'мразь',
'взял': 'взял',
'человека': 'человека',
'нахуй': 'нахуй',
'креативно': 'креативно',
'психология': 'психология',
'примерно': 'примерно',
'терминал': 'терминал',
'этих': 'oldтих',
'костылей': 'костыль',
'должен': 'должне',
```

```
'этого': 'этого',
'terminal': 'terminal',
'multiplexer': 'multiplexer',
'фотошоп': 'фотошоп',
'научиться': 'научитися',
'использовать': 'использовати',
'много': 'много',
'ума': 'умо',
'ломанули': 'лануеи',
'slavmetal': 'slavmetal',
'termite': 'termite',
'умеет': 'уметь',
'вкладки': 'вкладок',
'нееееееее': 'нееееe',
'винду': 'виндаджа',
'ламер': 'ламерамламела',
'извините': 'извинити',
'куда': 'oldуда',
'деваться': 'деваться',
'ааа': 'ааа',
'видать': 'видити',
'права': 'право',
'необходимые': 'необходим',
'выставить': 'выставить',
'sudo': 'sudo',
'chmod': 'chmod',
'захожу': 'захоги',
'через': 'через',
'ритуал': 'ритуал',
'религии': 'религия',
'перед': 'перед',
'каждой': 'каждий',
'коммандой': 'кманде',
'молимся': 'молиться',
'супер': 'супер',
'пользователю': 'пользователь',
'каждую': 'каждувати',
'ебалу': 'ебалу',
'судо': 'oldудо',
'запускаешь': 'запускать',
'честно': 'честно',
'конфиг': 'конфl',
'home': 'home',
'щас': 'щас',
'серьезно': 'серьезно',
'оба': 'old',
'поняли': 'поняти',
'автостарт': 'аостае',
'сессию': 'сессия',
'перезапускать': 'перезапускать',
'запуск': 'запуск',
'привычка': 'привычка',
'кофиг': 'кофиг',
'локальный': 'локальный',
'аааааа': 'аааааа',
'вырезал': 'везал',
'часть': 'часть',
'конфига': 'конфига',
'абсолютно': 'абсолютно',
'меняется': 'меняется',
```

```
'выполняю': 'выполнять',
'хотя': 'oldотя',
'передёрнуть': 'передёрнуть',
'пытаюсь': 'пртатися',
'изучить': 'изучати',
'напомни': 'напомнить',
'вывести': 'вовести',
'них': 'old',
'хотел': 'хотел',
'лету': 'лёт',
'хочешь': 'хочel',
'コニキはシテトトイレレヴロソフトです': 'sニキはシテトトイレレヴロソフトです',
'хоть': 'oldоть',
'меняю': 'меняйяняйня',
'ничего': 'ничel',
'про': 'old',
'каво': 'кавококо',
'фройляйн': 'фйляен',
'chen': 'chen',
'lein': 'lein',
'уменьшительно': 'уменьшительно',
'ласкательное': 'ласкательное',
'германских': 'германские',
'языках': 'язык',
'спасибо': 'ссибо',
'config': 'config',
'первую': 'первувати',
'очередь': 'очередь',
'название': 'названин',
'бесит': 'беситисит',
'wiki.archlinux.org': 'wiki.archlinux.org',
'редактируй': 'редактируе',
'свои': 'oldвои',
'конки': 'конки',
'выполнил': 'выполнить',
'палец': 'палец',
'устал': 'устал',
'скроллить': 'соллие',
'мобиле': 'мобил',
'телегу': 'телега',
'влезает': 'влезать',
'давайте': 'давати',
'длинные': 'длиннно',
'рукописные': 'рукописные',
'тексты': 'тексты',
'пасту': 'паст',
'кидать': 'кидити',
'паста': 'пасто',
'телеграф': 'телеграф',
'mkdir': 'mkdir',
'print': 'print',
'говно': 'говновно',
'офигел': 'офигел',
'вываливать': 'вываливате',
'удобно': 'удобно',
'system': 'system',
'monitor': 'monitor',
'based': 'based',
'torsmo': 'torsmo',
'original': 'original',
```

```
'code': 'code',
'licensed': 'license',
'bsd': 'bsd',
'license': 'license',
'written': 'written',
'fork': 'fork',
'gpl': 'gpl',
'please': 'please',
'see': 'see',
'copying': 'copying',
'details': 'details',
'copyright': 'copyright',
'hannu': 'hannu',
'saransaari': 'saransaari',
'lauri': 'lauri',
'hakkarainen': 'hakkarainen',
'brenden': 'brenden',
'matthews': 'matthew',
'philip': 'philip',
'kovacs': 'kovac',
'authors': 'authors',
'rights': 'rights',
'program': 'program',
'free': 'free',
'software': 'software',
'can': 'can',
'redistribute': 'redistribute',
'modify': 'modify',
'terms': 'terms',
'gnu': 'gnu',
'general': 'general',
'public': 'public',
'published': 'publish',
'foundation': 'foundation',
'either': 'either',
'version': 'version',
'option': 'option',
'later': 'later',
'distributed': 'distributed',
'hope': 'hope',
'will': 'will',
'useful': 'useful',
'without': 'without',
'warranty': 'warranty',
'even': 'even',
'implied': 'implied',
'merchantability': 'merchantability',
'fitness': 'fitness',
'particular': 'particular',
'purpose': 'purpose',
'received': 'received',
'copy': 'copy',
'along': 'along',
'www': 'www',
'org': 'org',
'licenses': 'license',
'alignment': 'alignment',
'top': 'top',
'left': 'left',
'background': 'background',
```

```
'false': 'false',
'border': 'border',
'width': 'width',
'cpu': 'cpu',
'avg': 'avg',
'samples': 'samples',
'default': 'default',
'color': 'color',
'white': 'white',
'outline': 'outline',
'shade': 'shade',
'double': 'double',
'buffer': 'buffer',
'true': 'true',
'draw': 'draw',
'borders': 'border',
'graph': 'graph',
'shades': 'shades',
'extra': 'extra',
'newline': 'newline',
'font': 'font',
'dejavu': 'dejavu',
'sans': 'sans',
'mono': 'mono',
'gap': 'gap',
'minimum': 'minimum',
'height': 'height',
'net': 'net',
'buffers': 'buffer',
'console': 'console',
'ncurses': 'ncurse',
'stderr': 'stderr',
'window': 'window',
'class': 'class',
'type': 'type',
'show': 'show',
'range': 'range',
'scale': 'scale',
'stippled': 'stipple',
'interval': 'interval',
'uppercase': 'uppercase',
'use': 'use',
'spacer': 'spacer',
'none': 'none',
'xft': 'xft',
'text': 'text',
'grey': 'greyrey',
'info': 'info',
'scroll': 'scroll',
'sysname': 'sysname',
'nodename': 'nodename',
'kernel': 'kernel',
'machine': 'machine',
'uptime': 'uptime',
'frequency': 'frequency',
'mhz': 'mhz',
'freq': 'freq',
'ghz': 'ghz',
'usage': 'usagele',
'memmax': 'memmax',
```

```
'memperc': 'memperc',
'membar': 'membar',
'swap': 'swapae',
'swapmax': 'swapmax',
'swapperc': 'swapperc',
'swapbar': 'swapbar',
'cpubar': 'cpubar',
'processes': 'processes',
'running': 'running',
'file': 'file',
'systems': 'systems',
'used': 'used',
'bar': 'bar',
'networking': 'networking',
'upspeed': 'upspeed',
'downspeed': 'downspeed',
'name': 'name',
'pid': 'pid',
'lightgrey': 'lightgrey',
'конфигурация': 'конфигурация',
'коньков': 'кьков',
'английской': 'английский',
'вики': 'вика',
'посмотри': 'посмотр',
'себя': 'oldебя',
'выдаст': 'выдать',
'звук': 'звук',
'где': 'old',
'applet': 'applet',
'под': 'old',
'трей': 'oldрей',
'имею': 'имя',
'виду': 'вид',
'имеешь': 'иметь',
'ввиду': 'ввид',
'виджеты': 'виджет',
'панеле': 'панel',
'плазмоид': 'пзмоед',
'коробки': 'коробка',
'чтобы': 'чтобы',
'системный': 'системній',
'зачем': 'зачем',
'тебе': 'oldебе',
'плазмоиды': 'пзмоие',
'лол': 'лол',
'пока': 'oldока',
'сижу': 'сижтти',
'кедах': 'кед',
'думаю': 'думати',
'переходить': 'переходити',
'покажи': 'покажи',
'awesomewm': 'awesomewm',
'патчить': 'пчить',
'suckless': 'suckless',
'звучит': 'звучити',
'гораздо': 'гораздо',
'страшнее': 'страшный',
'самом': 'самом',
'деле': 'деле',
'посоветуешь': 'посоветуешь',
```

```
'всм': 'всм',
'дичь': 'дича',
'какая': 'какая',
'приоритете': 'приоритете',
'сколько': 'слько',
'ставить': 'ставити',
'равно': 'равно',
'открывает': 'открнвати',
'сстемный': 'семней',
'проги': 'проги',
'генерирую': 'герире',
'особенно': 'обенео',
'сделай': 'сделай',
'нету': 'нета',
'говорят': 'говорят',
'xterm': 'xterm',
'uses': 'use',
'located': 'located',
'понимаю': 'понимати',
'инструкциях': 'инструкция',
'пишется': 'писаться',
'находится': 'находитися',
'директории': 'директория',
'вводе': 'ввод',
'знаете': 'знати',
'жопу': 'жоп',
'смазать': 'сзать',
'жопой': 'жопа',
'ешь': 'есть',
'девать': 'девal',
'80мб': '80мб',
'окно': 'окно',
'жесть': 'жесть',
'kitty': 'kitty',
'попробовал': 'попробовать',
'моргает': 'моргать',
'белым': 'белые',
'терминала': 'терминал',
'невыолнимом': 'невыолнимом',
'инпуте': 'инпуl',
'клавы': 'клавы',
'конец': 'конец',
'файла': 'файла',
'тыкаю': 'тыкаю',
'клавишу': 'клавиш',
'вниз': 'oldниз',
'жмешь': 'жмешь',
'esc': 'esc',
'отключить': 'отключить',
'конфиге': 'конфиге',
'нашел': 'нашел.',
'zachrootишься': 'zachrootишься',
'потом': 'потом',
'сломаешь': 'смаееь',
'всегда': 'всегl',
'почему': 'почel',
'думал': 'думал',
'один': 'olddин',
'аккаунт': 'аккаунт',
'пускает': 'пускать',
```

```
'arch': 'arch',
'forum': 'forum',
'аккаунты': 'аккаунт',
'разные': 'разний',
'залипает': 'залипает',
'шелезяке': 'шезяее',
'достаточно': 'достаточно',
'клаву': 'клавав',
'ноут': 'ноута',
'костылять': 'ктыляе',
'заебёшься': 'збёшье',
'имо': 'old',
'свитч': 'свитч',
'свисток': 'свистка',
'могли': 'могти',
'сломать': 'сломам',
'драйвер': 'драйвер',
'ядре': 'ядро',
'можешь': 'можити',
'сам': 'old',
'собрать': 'собраття',
'юзать': 'юзатити',
'dkms': 'dkm',
'xset': 'xset',
'rate': 'rate',
'помогает': 'помогает',
'пофиксить': 'пофиксити',
'дребезг': 'дбезг',
'клавиш': 'клавиша',
'клава': 'клавав',
'дохнет': 'дохнl',
'али': 'алиалаля',
'возможно': 'возможно',
'запилили': 'запилити',
'лтс': 'old',
'теперь': 'тепеl',
'кажется': 'кется',
'дилема': 'дилема',
'звуковуха': 'зковуе',
'плохо': 'плохохохомохохохохохохохохохохохохохохохохом',
'lts': 'lt',
'wifi': 'wifi',
'пашет': 'пашетат',
'ядрами': 'ядро',
'замкнутый': 'замкнутый',
'круг': 'круг',
'обновлять': 'обновлять',
'рано': 'oldано',
'поздно': 'поздно',
'навернётся': 'навернётсе',
'разбирается': 'разбираться',
'хочу': 'хотіти',
'выше': 'выше',
'устраивает': 'устраивать',
'внести': 'внестистестисти',
'игнор': 'игнор',
'пакетов': 'пакет',
'linux': 'linux',
'headers': 'headers',
'api': 'api',
```

```
 'firmware': 'firmware',
 'libutil': 'libutil',
 'glibc': 'glibc',
 'gcc': 'gcc',
 'libs': 'lib',
 'ещё': 'old',
 'какие': 'какие',
 'закидывать': 'закидывать',
 'исключения': 'исключение',
 'плане': 'пласти',
 'рабочее': 'рабочий',
 'окружение': 'окружение',
 'настраивать': 'настраивати',
 'бекап': 'бекап',
 'впадлу': 'впадвдсдаддаддаддаддададдаддаддаддаддаддаддаддаддаддаддаддаддад
даддаддаддаддаддаддаддаддаддаддаддаддаддаддаддаддаддвддаддаддадда',
 'трус': 'трус',
 'ага': 'ага',
 'уже': 'old',
 'настроенный': 'настроить',
 'граб': 'oldраб',
 'пытаться': 'пытитися',
 'махнуть': 'мнуть',
 'рефинд': 'рефил',
 'путь': 'путь',
 'выстрелу': 'втрееу',
 'колено': 'колел',
 'отсутствие': 'отсутствие',
 'графического': 'графический',
 'установщика': 'установщика',
 'создает': 'создавать',
 'рабочие': 'рабочий',
 'места': 'место',
 'gucharmap': 'gucharmap',
 'вспомнил': 'вспомнити',
 'ребят': 'ребото',
 'напомните': 'напомните',
 'плиз': 'плизизсизкз',
 'называется': 'назавается',
 'софтина': 'стина',
 'которой': 'корой',
 'просматривать': 'просматривати',
 'глифы': 'глифы',
 'бэкапы': 'бэкап',
 'рулят': 'рулить',
 'следующий': 'следующий',
 'обделаешься': 'обделаешься',
 'большому': 'большой',
 'обделался': 'оелале',
 'лёгким': 'лёгкое',
 'испугом': 'иугом',
 'восстановил': 'восстановил',
 'эти': 'old',
 'файлы': 'файлы',
 'бэкапа': 'бэкal',
 'прошло': 'прости',
 'пакеты': 'пакеты',
 'конфликтуют': 'конфликтовать',
 'причина': 'причина',
 'установлена': 'установить',
```

```
 'время': 'время',
 'последнего': 'последне',
 'падения': 'падение',
 'восстановления': 'восстановление',
 'fsck': 'fsck',
 'повредились': 'повредились',
 'var': 'var',
 'lib': 'lib',
 'пакет': 'пакет',
 'files': 'files',
 'overwrite': 'overwrite',
 'буду': 'oldуду',
 'пробовать': 'пробовать',
 'force': 'forceercecrcecrescroreocoreoccolencccccecceccecceccccccccccccc
ccccccccccccccccccrecccccccccccrecrecceccecceccecceccecceccecceccecceccecceccecce
ccecсec',
 'нагуглил': 'нуглел',
 'параметр': 'параметр',
 'распознаётся': 'распознаётся',
 'например': 'например',
 'понял': 'понял',
 'мож': 'old',
 'перезагрузить': 'перезагрузить',
 'комп': 'комп',
 'короткая': 'коротка',
 'помню': 'помня',
 'обновляю': 'обновлять',
 'систему': 'система',
 'suy': 'suy',
 'ставишь': 'ставить',
 'должно': 'должl',
 'подсказывает': 'подсказывать',
 'чуйка': 'чуйка',
 'овердохуя': 'овердохуя',
 'переименовывал': 'переименовывал',
 'коллеги': 'коллега',
 'пардон': 'пардондод',
 'тока': 'тока',
 'онлайн': 'онлайн',
 'юзал': 'юзалалалолюлазлюзлазлазлазлазлазлазлазлазлазлазлазлазлазлаз
лазлазлазлюзазлазлазлюзазлюзазлюзазлазюзазазлазазазлазлюза',
 'принимаюсь': 'принимаюсе',
 'иксы': 'oldксы',
 'еееe': 'ееейе',
 'сеть': 'сеть',
 'заработала': 'заработать',
 'словари': 'словарь',
 'вся': 'old',
 'хуйня': 'хуйній',
 'скачать': 'скачити',
 'словарь': 'словарь',
 'пишет': 'пишет',
 'недоступен': 'недоступен',
 'работала': 'работать',
 'чрутом': 'чрутl',
 'флешки': 'флешок',
 'костыль': 'костыль',
 'фурычала': 'фычаеа',
 'chroot': 'chroot',
 'livecd': 'livecd',
```

```
'установить': 'установить',
'спервa': 'спервa',
'мне': 'old',
'любом': 'люб',
'случае': 'случай',
'помучаться': 'помучаться',
'сетью': 'сеть',
'мучаться': 'матьея',
'pikaur': 'pikaur',
'окей': 'окойа',
'разбираться': 'разбиритися',
'ним': 'він',
'networkmanager': 'networkmanager',
'вешайте': 'вешать',
'понятно': 'понятно',
'вешали': 'вешal',
'знал': 'знал',
'сути': 'сути',
'копипастнул': 'копипастнул',
'арчвики': 'авики',
'хостнейм': 'хтнеем',
'хост': 'хост',
'установке': 'установка',
'делали': 'делал',
'хостс': 'хостс',
'чето': 'чето',
'ожидал': 'ожидать',
'такое': 'такое',
'произойдет': 'произойти',
'таки': 'таки',
'инет': 'oldнет',
'такс': 'такс',
'падажи': 'падal',
'забудь': 'забудити',
'рамас': 'рамас',
'аниме': 'аниме',
'девочек': 'девочка',
'рекламы': 'реклама',
'своих': 'своих',
'корыстных': 'кыстне',
'целей': 'цель',
'жестоко': 'жестоко',
'сидел': 'сидеть',
'исключительно': 'исключительно',
'дебиановских': 'дебиановских',
'перепривыкать': 'перепривыкать',
'тяжеловато': 'тяжеловатый',
'немного': 'нного',
'заинсталлить': 'заинсталлить',
'представляю': 'представляти',
'насколько': 'насколько',
'начну': 'начать',
'понимать': 'понимати',
'сделаю': 'сделать',
'ближайшее': 'ближний',
'хватит': 'хватитти',
'интима': 'интим',
'арчем': 'арч',
'спс': 'спс',
'инстол': 'инстl',
```

```
 'xfce': 'xfce',
 'наверное': 'наверно',
 'какой': 'какой',
 'будешь': 'будел',
 'никакого': 'никакий',
 'congratz': 'congratz',
 'молодец': 'молодец',
 'арч': 'арч',
 'удоли': 'удоли',
 'сори': 'сор',
 'чатом': 'чат',
 'ошибся': 'ошибиться',
 'гента': 'гента',
 'долго': 'долго',
 'компилится': 'компилитсе',
 'скажи': 'скажи',
 'три': 'old',
 'слова': 'слово',
 'которые': 'корые',
 'услышать': 'услышити',
 'ноль': 'ноль',
 'конфигурации': 'конфигурация',
 'копирую': 'кирую',
 'live': 'live',
 'hosts': 'host',
 'network': 'network',
 'иксах': 'иксах',
 'кратко': 'кратко',
 'том': 'old',
 'ставлю': 'ставля',
 'плазме': 'плазма',
 'управлять': 'управити',
 'звуком': 'звук',
 'альсы': 'альсы',
 'аплета': 'аплел',
 'крайняк': 'крайняк',
 'ладно': 'ладно',
 'добавить': 'добавити',
 'будто': 'будто',
 'статический': 'статический',
 'крайнем': 'крайний',
 'работать': 'работити',
 'верно': 'верно',
 'шаг': 'шаг',
 ...}
```

0.4 Messages

In [8]:
```
df = pd.read_csv(DIALOGS_MERGED_DATA_PATH)
df['date_time'] = pd.to_datetime(df['date'])
df['date'] = df['date_time'].dt.date
df['from_id'] = df['from_id'].str.extract(r'(\d+)').fillna(0).astype(int)
df['to_id'] = df['to_id'].str.extract(r'(\d+)').fillna(0).astype(int)

df_meta = pd.read_csv(DIALOGS_META_MERGED_DATA_PATH)
df = df.merge(df_meta.drop_duplicates(subset='dialog_id'), on='dialog_id'
df = df.rename(columns={'type_x': 'message_type', 'type_y': 'dialog_type'

def process_message(message):
```

```
    if not isinstance(message, str):
        return ""
    message = message.lower()
    # Remove –
    message = message.replace("–", "").replace("—", "")
    # Remove punctuation and other symbols
    message = re.sub(f"[{re.escape(string.punctuation)}]", "", message)
    # Remove numbers and extra spaces
    message = re.sub(r"\d+", "", message).strip()
    # Tokenize, remove stopwords, and join back into a string
    return " ".join([word for word in message.split() if word.lower() not

def add_tones(words):
    words = df['message_proc'].str.split()
    words_lemmatized = words.apply(lambda words: [dict_lemmatized.get(w,
    df['tones'] = words_lemmatized.apply(lambda words: [tone_dict.get(w,
    df['tone'] = df['tones'].apply(lambda tones: sum(tones))
    # Create normalized score accross identified words (words with tone !
    df['tone_normalized'] = df['tones'].apply(lambda tones: sum(tones) /
    return df

def compute_document_embedding(words, model):
    valid_words = [word for word in words if word in model.wv]  # Keep on
    if not valid_words:
        return np.zeros(model.vector_size)  # Return a zero vector if no
    return np.mean([model.wv[word] for word in valid_words], axis=0)

df['message_proc'] = df['message'].apply(process_message)
df = add_tones(df['message_proc'])

corpus = df['message_proc'].dropna().str.split().tolist()
w2v_model = Word2Vec(sentences=corpus, vector_size=128, window=5, min_cou
df['word2vec'] = df['message_proc'].str.split().apply(lambda x: compute_d
# Normalize word2vec embeddings
df['word2vec_norm'] = df['word2vec'].apply(lambda x: x / np.linalg.norm(x
```

In [9]:
```
df.info()
```

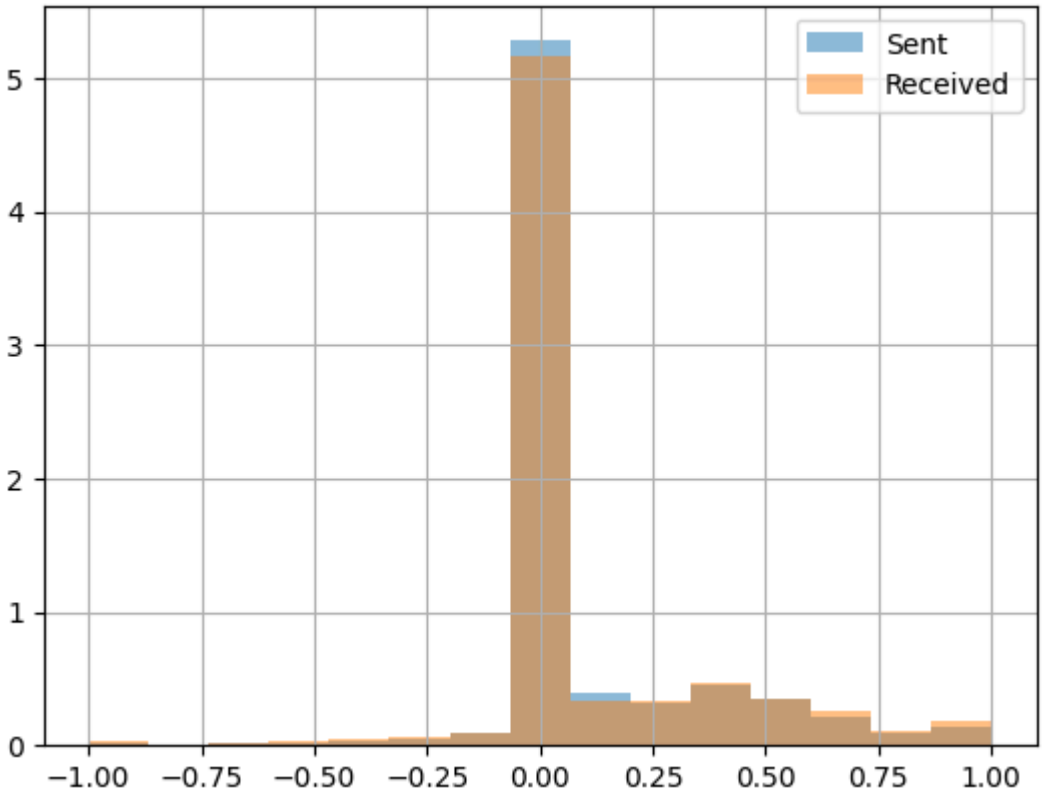
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 315271 entries, 0 to 315270
Data columns (total 20 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   id               315271 non-null  float64
 1   date             315271 non-null  object
 2   from_id          315271 non-null  int64
 3   to_id            315271 non-null  int64
 4   fwd_from         10303 non-null   object
 5   message          273248 non-null  object
 6   message_type     315271 non-null  object
 7   duration         7627 non-null    float64
 8   reactions        280746 non-null  object
 9   dialog_id        315271 non-null  int64
 10  date_time        315271 non-null  datetime64[ns, UTC]
 11  name             241146 non-null  object
 12  dialog_type      241147 non-null  object
 13  users            241147 non-null  object
 14  message_proc     315271 non-null  object
 15  tones            315271 non-null  object
 16  tone             315271 non-null  float64
 17  tone_normalized  315271 non-null  float64
 18  word2vec         315271 non-null  object
 19  word2vec_norm    315271 non-null  object
dtypes: datetime64[ns, UTC](1), float64(4), int64(3), object(12)
memory usage: 48.1+ MB
```

1. Sentiment analysis

1.1 What is the distibution of tones of messages I've sent vs messages I've received?

In [10]:
```
send_df = df[df['from_id']==MY_UID]
recv_df = df[df['to_id']==MY_UID]
send_df.tone_normalized.hist(bins=15, alpha=0.5, density=True, label='Sen
recv_df.tone_normalized.hist(bins=15, alpha=0.5, density=True, label='Rec
plt.legend()
```

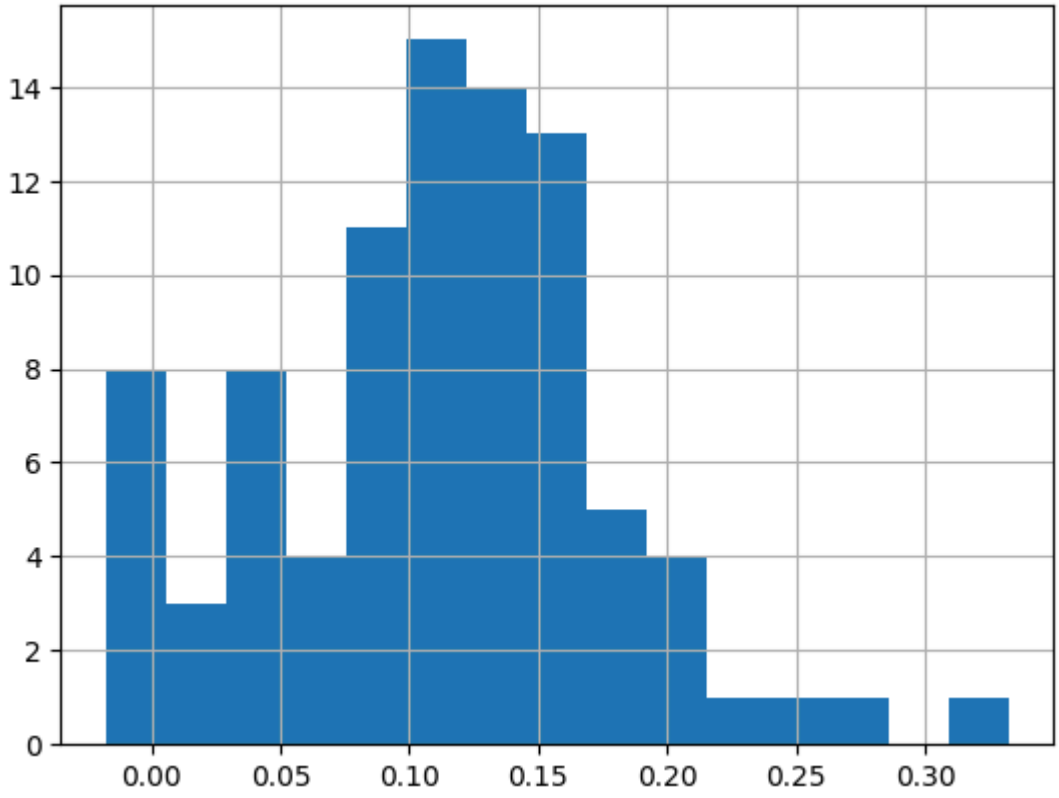
Out[10]:
```
<matplotlib.legend.Legend at 0x367290410>
```



1.2 What are the distribution of message tones accross groups?

In [11]:
```
df[df['dialog_type'] == 'Group'].groupby('dialog_id')['tone_normalized'].
```

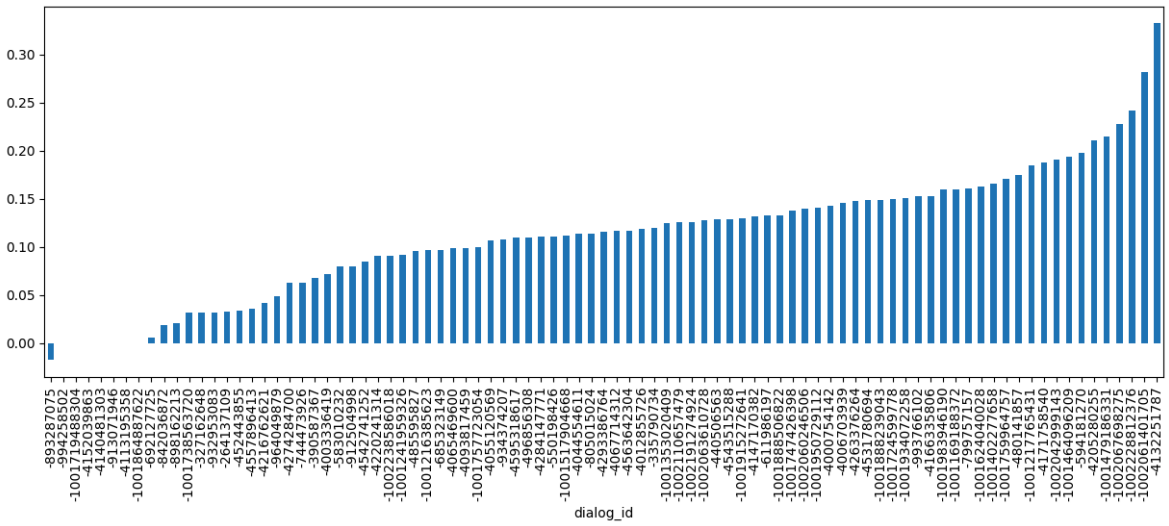
Out[11]: <Axes: >



1.3 What are the average tone of messages for different groups?

In [12]:
```
plt.figure(figsize=(15, 5))
df[df['dialog_type'] == 'Group'].groupby('dialog_id')['tone_normalized'].
```

Out[12]: <Axes: xlabel='dialog_id'>

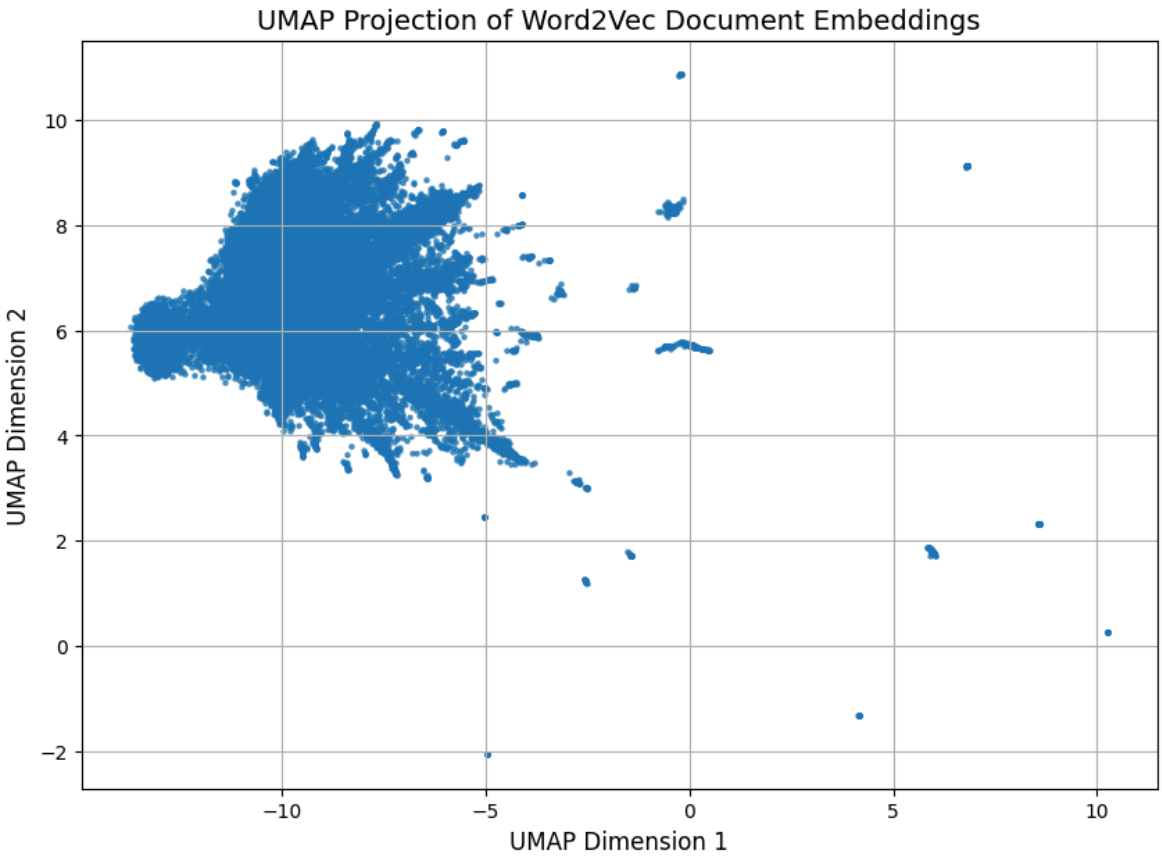


2. Topic Modeling

Here we aggregated the messages by dialog and day. Then performed HDBSCAN clusterization to identify the clusters (topics)

In [13]:
```
df_ = (
    df.groupby(['dialog_id', 'date'])
    .agg({
        'message_proc': lambda messages: ' '.join(messages),
        'tone': 'sum',
        'tone_normalized': 'mean',
        'word2vec': lambda vectors: np.mean(np.stack(vectors), axis=0),
        'from_id': 'count'
    })
    .reset_index()
    .rename(columns={'from_id': 'message_count'})
)
```

In [14]:
```
document_vectors = np.vstack(df_['word2vec'].values)
umap_reducer = umap.UMAP(n_neighbors=15, min_dist=0.1, metric='euclidean'
reduced_vectors_umap = umap_reducer.fit_transform(document_vectors)

plt.figure(figsize=(10, 7))
plt.scatter(reduced_vectors_umap[:, 0], reduced_vectors_umap[:, 1], s=5,
plt.title("UMAP Projection of Word2Vec Document Embeddings", fontsize=14)
plt.xlabel("UMAP Dimension 1", fontsize=12)
plt.ylabel("UMAP Dimension 2", fontsize=12)
plt.grid(True)
plt.show()
```



In [15]:
```
document_vectors = np.vstack(df_['word2vec'].values)

hdbscan_clusterer = hdbscan.HDBSCAN(
    min_cluster_size=10,
    min_samples=3,
    cluster_selection_epsilon=0.005,
    core_dist_n_jobs=-1
)
clusters = hdbscan_clusterer.fit_predict(document_vectors)

df_['cluster'] = clusters

cluster_common_words = {}
for cluster_id in sorted(df_['cluster'].unique()):
    if cluster_id == -1:
        continue
    cluster_messages = ' '.join(df_.loc[df_['cluster'] == cluster_id, 'me
    word_counts = Counter(cluster_messages.split())
    cluster_common_words[cluster_id] = word_counts.most_common(100)

for cluster_id, common_words in cluster_common_words.items():
    print(f"Cluster {cluster_id}:")
    for word, count in common_words:
        print(f"  {word}: {count}")
    print()
```

Cluster 0:
  перекличка: 12

Cluster 1:
  👋: 15

Cluster 2:
  жк: 44
  avalon: 44
  prime: 44
  хід: 43
  будівництва: 43
  будинок: 43
  новини: 17
  🔶оновили: 17
  дивитись: 17
  деталі: 17
  фото🔶: 16
  збудовано: 15
  лютий: 3
  березень: 3
  квітень: 3
  травень: 3
  червень: 3
  липень: 3
  серпень: 3
  жовтень: 2
  листопад: 2
  грудень: 2
  будується: 1
  характеристики🔶: 1
  поверховість: 1

Cluster 3:
  новим: 22
  роком: 22
  🥰🥰🥰: 9
  сонце😚❤: 1
  ✨✨✨: 1
  🎄🎄🎄: 1

Cluster 4:
  👋: 16
  митник: 3
  ставши: 3
  найбільший: 3
  набір: 2
  алгоритм: 2
  маємо: 2
  розглянути: 2
  якись: 2
  фарисей: 2
  боже: 2
  царстві: 2
  покликав: 2
  бувши: 2
  смерти: 2
  дорозі: 2
  знайшов: 1
  відео: 1

розбирати: 1
штука: 1
айфонах: 1
працює: 1
🤓: 1
опис: 1
написав: 1
ростислав: 1
схоже: 1
даних: 1
результатів: 1
правильно: 1
зрозумів: 1
обрати: 1
проблему: 1
вирішує: 1
разом: 1
придумати: 1
задачу: 1
вирішувати: 1
розумію: 1
вибір: 1
вільний: 1
теми: 1
розгляді: 1
думаю: 1
цікаво: 1
реалізувати: 1
тісно: 1
звязане: 1
ла: 1
допомоги: 1
сторонніх: 1
бібліотек: 1
numpy: 1
звісно: 1
❤: 1
привіт: 1
тонкое: 1
искусство: 1
пофигизма: 1
парадоксальный: 1
способ: 1
жить: 1
счастливо: 1
ммэнсон: 1
httpsplaygooglecomstorebooksdetailsidledwaaqbaj: 1
«а: 1
певні: 1
праведні: 1
мали: 1
притчу: 1
розповів: 1
чоловіки: 1
храму: 1
ввійшли: 1
помолитись: 1
молився: 1
здирщики: 1
неправедні: 1

```
  перелюбні: 1
  пощу: 1
  рази: 1
  тиждень: 1
  даю: 1
  десятину: 1
  надбаю: 1
  здалека: 1
  стояв: 1
  очей: 1
  звести: 1
  неба: 1
  смів: 1
  бив: 1
  груди: 1
  казав: 1
  милостивий: 1
  грішного: 1
  говорю: 1
  повернувся: 1
  дому: 1
  виправданий: 1

Cluster 5:
  remote: 653
  data: 397
  scientist: 385
  engineer: 282
  experience: 145
  senior: 78
  kyiv: 46
  learning: 43
  machine: 37
  junior: 36
  luxoft: 35
  дайджест: 32
  привіт: 31
  вакансій👇: 31
  лови: 30
  свіжий: 30
  nlp: 28
  lead: 23
  alistware: 22
  ds: 22
  dl: 19
  middle: 16
  software: 15
  itrs: 15
  nix: 14
  science: 12
  kharkiv: 12
  samsung: 12
  intellias: 12
  lviv: 11
  globallogic: 11
  middlesenior: 11
  simporter: 9
  exadel: 9
  developer: 8
  solutions: 8
```

metinvest: 8
tech: 8
ciklum: 7
приватбанк: 7
vision: 7
infopulse: 7
specialist: 6
aiml: 6
capital: 6
recruiters: 6
fozzy: 6
вакансий: 6
deep: 6
sciforce: 6
rozetka: 6
ipland: 6
team: 5
softconstruct: 5
relocate: 5
litslink: 5
cvml: 5
mlai: 5
geocomply: 5
researcher: 5
analytics: 5
chi: 5
softserve: 5
zoral: 5
pheon: 5
mira: 5
lab: 5
привет: 4
свежий: 4
shelf: 4
mlcv: 4
soft: 4
dataart: 4
dnipro: 4
psr: 4
autodoc: 4
unikoom: 4
ncube: 4
eleks: 4
itjim: 4
winstars: 4
architect: 4
spdukraine: 4
global: 4
wix: 4
softum: 4
svitla: 4
bank: 4
лун: 4
novos: 4
trainee: 3
itexpert: 3
ловите: 3
depdiko: 3
relocation: 3
mgid: 3

technology: 3
grid: 3
dynamics: 3
privatbank: 3

Cluster 6:
України: 4977
✅: 3137
стартапів: 3049
Україні: 2826
👉: 2757
має: 2658
цифрової: 2636
розвитку: 2550
підтримки: 2269
питання: 2268
youtube: 2247
шо: 2206
трансформації: 2093
посиланням: 2030
зробити: 2028
участь: 2012
даних: 1971
працює: 1935
разом: 1923
модель: 1919
отримати: 1912
навчання: 1899
грн: 1882
думаю: 1808
можливість: 1800
онлайн: 1797
програми: 1748
цікаво: 1731
роботи: 1720
людей: 1698
вишкіл: 1688
робити: 1674
клуб: 1606
проєкту: 1605
закритий: 1572
дія: 1553
взагалі: 1522
типу: 1514
відео: 1511
привіт: 1508
компанії: 1478
бізнесу: 1461
рішення: 1410
українських: 1400
data: 1399
послуг: 1383
дії: 1366
□: 1340
можливо: 1335
знаю: 1309
ок: 1309
дані: 1297
студентів: 1281
📌: 1277

  мінцифри: 1272
  точно: 1266
  проєкт: 1265
  🔷: 1252
  👉🏻: 1248
  сфері: 1239
  мають: 1234
  нові: 1218
  наприклад: 1216
  курс: 1215
  послуги: 1215
  моделі: 1198
  завтра: 1182
  українців: 1172
  день: 1169
  ▪️: 1139
  життя: 1069
  команди: 1062
  млн: 1047
  можливості: 1035
  роботу: 1034
  речі: 1025
  знайти: 1019
  працювати: 1016
  війни: 1014
  фонду: 1012
  кількість: 997
  зможуть: 994
  розумію: 989
  допомогою: 974
  місце: 968
  відбудеться: 962
  реєстрація: 960
  компаній: 956
  української: 948
  взяти: 937
  завдання: 934
  стартапи: 931
  досвід: 912
  нема: 908
  новий: 907
  🔶: 900
  спільно: 892
  участі: 892
  освіти: 891
  міністерства: 890

Cluster 7:
  →: 1208
  data: 615
  scientist: 558
  engineer: 488
  experience: 295
  ⏩: 294
  learning: 125
  dataroot: 106
  machine: 98
  привіт: 87
  дайджест: 87
  потрібна: 87

допомога: 87
підготовці: 87
співбесіди: 87
команда: 87
university: 87
готова: 87
допомогти: 87
талановитим: 87
фахівцям: 87
зв’язок: 87
лови: 86
вакансій👇: 86
свіжий: 85
senior: 77
lead: 66
twitter: 64
linkedin: 64
datarootlabs: 64
🔥: 62
software: 53
створенні: 49
nlp: 48
vision: 44
створені: 38
mlops: 37
deep: 37
science: 37
rozetka: 36
developer: 36
labs: 32
globallogic: 30
dataforest: 28
🤙: 23
specialist: 21
quantum: 21
scalarr: 19
🤖: 19
artellence: 18
sigma: 18
adaptiq: 17
tech: 16
researcher: 15
team: 15
systems: 15
aiml: 15
technology: 14
alistware: 14
🇺🇦: 13
litslink: 13
chi: 13
prompt: 13
ukraine: 12
teamdev: 12
competera: 12
coxit: 12
asoft: 12
autodoc: 12
privatbank: 12
center: 12
megogo: 12

```
sevenpro: 11
generative: 11
knowledge: 11
junior: 10
пумб: 10
epicentrk: 10
acceptic: 10
devico: 10
beter: 10
winstars: 9
yael: 9
intern: 9
eastern: 9
peak: 9
llm: 9
jooble: 8
fozzy: 8
монвел: 8
uvik: 8
otakoyi: 8
zoral: 7
incoalliance: 7
codecare: 7
digital: 7
architect: 7
bank: 7
ajax: 7
adimen: 7
```

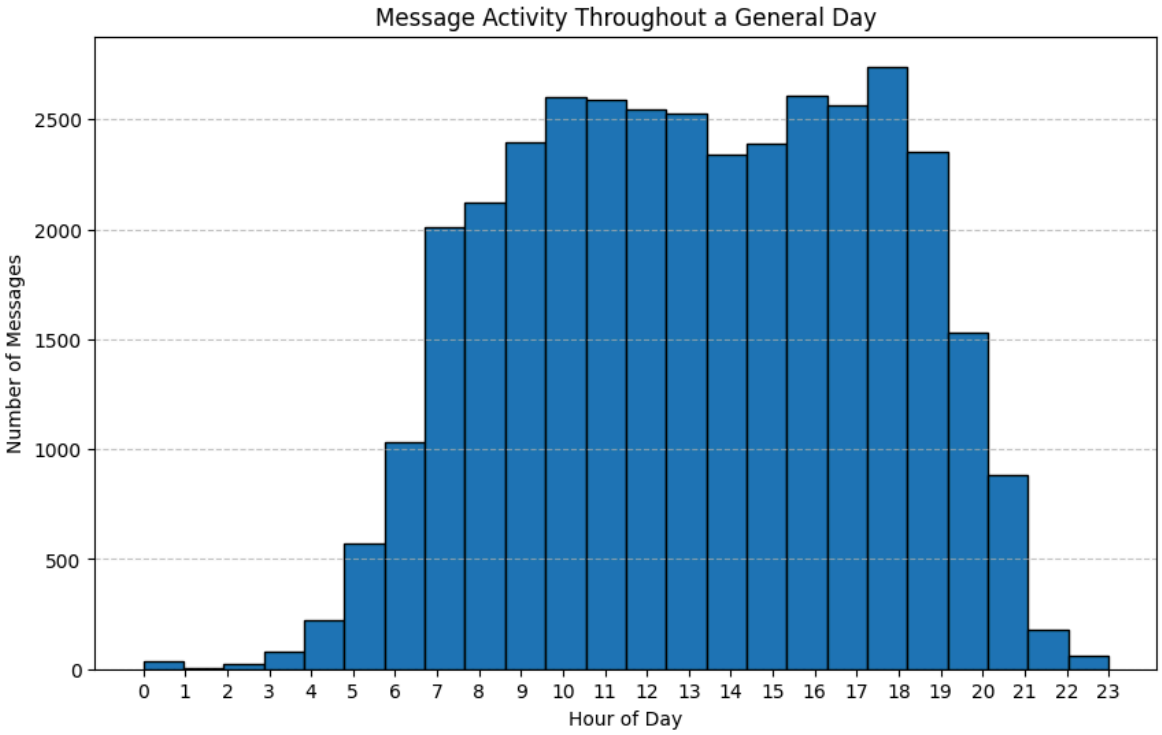
- Cluster 0 and 1: Looks like outliers
- Cluster 2: Is the topic of real estate announcements
- Cluster 3: Is the greetings with New Year
- Cluster 4: Is a mix of many topics
- Cluster 5: Is the topic of IT job offerings
- Cluster 6: Is the topic of Ukrainian startups
- Cluster 7: Is another topic of IT job offerings

3. Messages activity

3.1. What is my activity during the day?

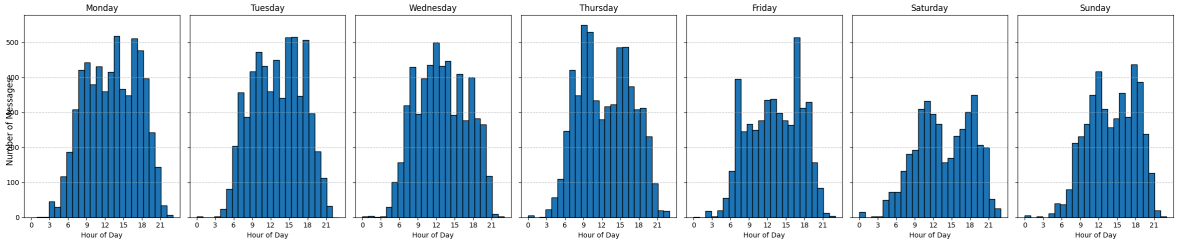
In [16]:
```
sent_df = df[df['from_id'] == MY_UID].copy()
sent_df['day_of_week'] = sent_df['date_time'].dt.day_name()
sent_df['hour'] = sent_df['date_time'].dt.hour

plt.figure(figsize=(10, 6))
plt.hist(sent_df['hour'], bins=24, range=(0, 23), edgecolor='black')
plt.title('Message Activity Throughout a General Day')
plt.xlabel('Hour of Day')
plt.ylabel('Number of Messages')
plt.xticks(range(0, 24))
```

```
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



3.2 What is my activity for each day of week?

In [17]:
```
sent_df = df[df['from_id'] == MY_UID].copy()
sent_df['day_of_week'] = sent_df['date_time'].dt.day_name()
sent_df['hour'] = sent_df['date_time'].dt.hour

day_order = ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Sat

fig, axes = plt.subplots(1, 7, figsize=(24, 5), sharey=True)

for ax, day in zip(axes, day_order):
    day_data = sent_df[sent_df['day_of_week'] == day]['hour']
    ax.hist(day_data, bins=24, range=(0, 23), edgecolor='black')
    ax.set_title(day)
    ax.set_xlabel('Hour of Day')
    ax.set_xticks(range(0, 24, 3))
    ax.grid(axis='y', linestyle='--', alpha=0.7)

fig.supylabel('Number of Messages', fontsize=12)
plt.tight_layout()
plt.show()
```

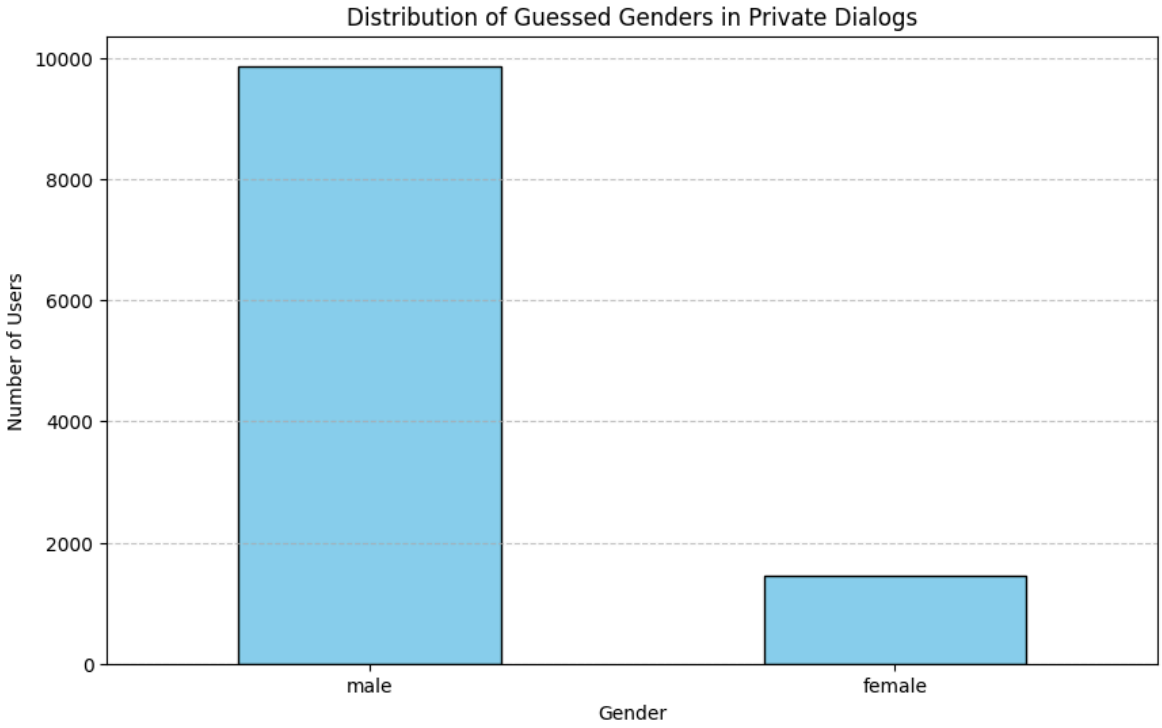


4. Recognize the interlocutors' genders

Here we used `gender_guesser` library to create rough ground truth. Then we have thrown out data points of unidentified genders and balanced classes by undersampling (oversampling didn't work well). Then we trained the Random Forest Classifier to predict gender based on message text.

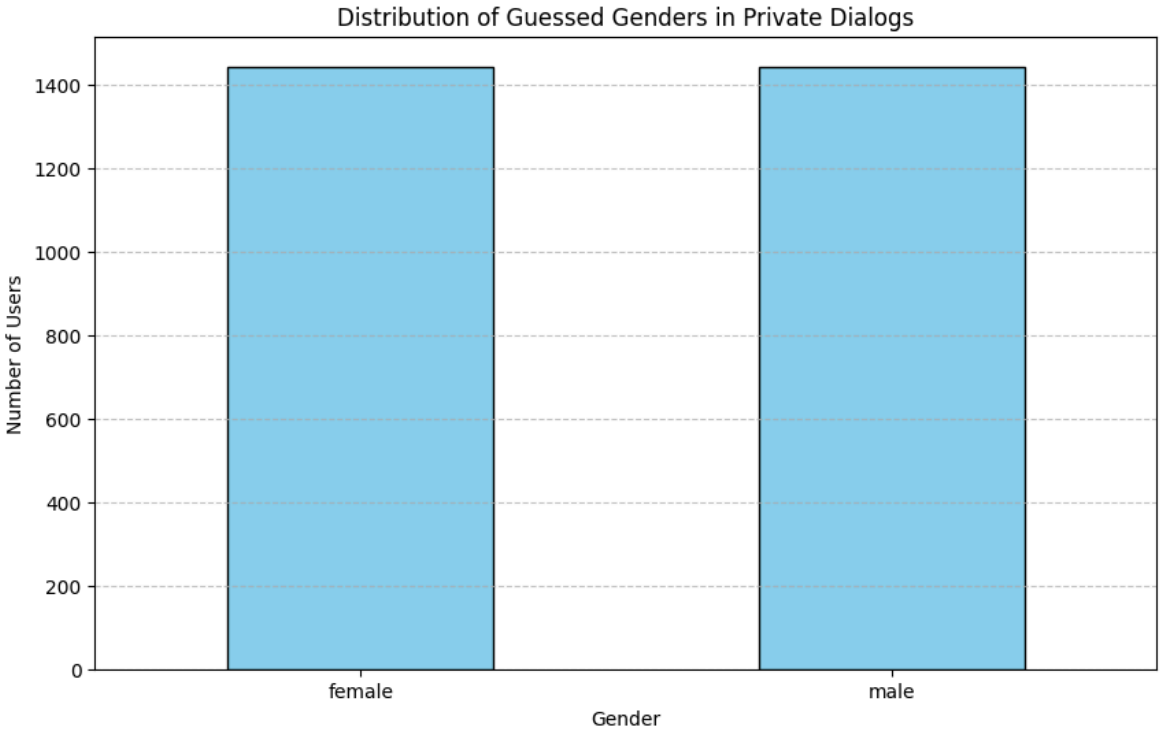
In [18]:
```
d = gender.Detector()
df_ = df.copy()
df_ = df_[df_['from_id'] != MY_UID]
df_ = df[df['dialog_type'] == 'Private dialog']
df_ = df_[df_['name'].isna() == False]

def guess_gender(full_name):
    name_parts = full_name.split()
    genders = [d.get_gender(part) for part in name_parts]
    normalized_genders = [g.replace('mostly_female', 'female').replace('m
    if not normalized_genders:
        return 'unknown'
    most_common_gender = Counter(normalized_genders).most_common(1)
    return most_common_gender[0][0] if most_common_gender else 'unknown'

df_['guessed_gender'] = df_['name'].apply(guess_gender)
df_ = df_[df_['guessed_gender'] != 'unknown']
df_ = df_[df_['message'].str.len() > 0]
```

In [19]:
```
gender_counts = df_['guessed_gender'].value_counts()
gender_counts.plot(kind='bar', color='skyblue', edgecolor='black', figsiz
plt.title('Distribution of Guessed Genders in Private Dialogs')
plt.xlabel('Gender')
plt.ylabel('Number of Users')
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



In [20]:
```
corpus = df_['message'].str.split().tolist()
w2v_model = Word2Vec(sentences=corpus, vector_size=128, window=5, min_cou
df_['word2vec_vector'] = df_['message'].str.split().apply(lambda x: compu
```

In [21]:
```
X = np.vstack(df_['word2vec_vector'].values)
y = df_['guessed_gender']

sampler = RandomUnderSampler(sampling_strategy=1, random_state=42)
X, y = sampler.fit_resample(X, y)

gender_counts = y.value_counts()
gender_counts.plot(kind='bar', color='skyblue', edgecolor='black', figsiz
plt.title('Distribution of Guessed Genders in Private Dialogs')
plt.xlabel('Gender')
plt.ylabel('Number of Users')
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



In [22]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,

model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred)
```

```
Accuracy: 0.5622837370242214

Classification Report:
              precision    recall  f1-score   support

      female       0.57      0.49      0.53       289
        male       0.55      0.63      0.59       289

    accuracy                           0.56       578
   macro avg       0.56      0.56      0.56       578
weighted avg       0.56      0.56      0.56       578
```

In [23]:
```
new_message = "Сонце!".split()
new_message_embedding = compute_document_embedding(new_message, w2v_model
predicted_gender = model.predict(new_message_embedding)
print("Predicted Gender:", predicted_gender[0])
```

```
Predicted Gender: female
```

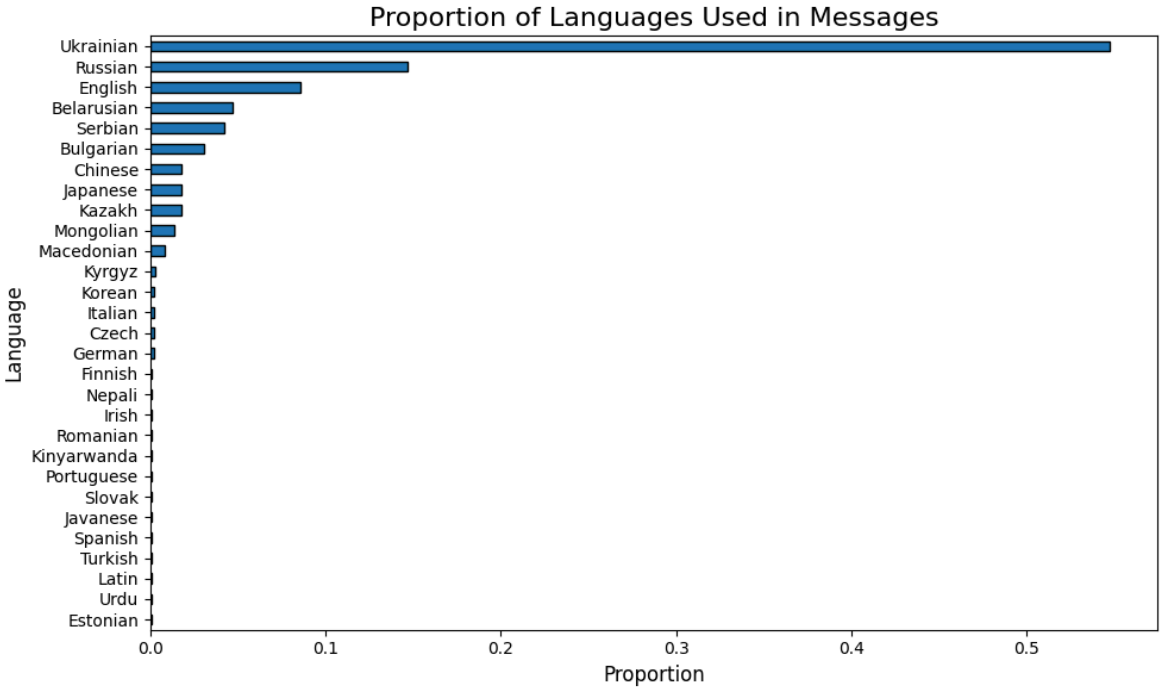
In [24]:
```
new_message = "Привіт!".split()
new_message_embedding = compute_document_embedding(new_message, w2v_model
predicted_gender = model.predict(new_message_embedding)
print("Predicted Gender:", predicted_gender[0])
```

```
Predicted Gender: male
```

5. What is the distribution of languages?

In [25]:

```
identifier = LanguageIdentifier.from_modelstring(lang_model, norm_probs=T
df_ = df[['message']].dropna().sample(1000).copy()

def detect_language(text):
    try:
        lang, prob = identifier.classify(text)
        return lang
    except Exception:
        return "unknown"

df_['language'] = df_['message'].apply(detect_language)

threshold = 0.001
language_counts = df_['language'].value_counts(normalize=True)
frequent_languages = language_counts[language_counts >= threshold]
other_languages = language_counts[language_counts < threshold].sum()

if other_languages > 0:
    frequent_languages['Others'] = other_languages

def get_language_name(code):
    try:
        return langcodes.get(code).language_name()
    except LookupError:
        return code

frequent_languages.index = [get_language_name(lang) for lang in frequent_

plt.figure(figsize=(10, 6))
frequent_languages.sort_values().plot(
    kind='barh',
    colormap='tab10',
    edgecolor='black'
)
plt.title('Proportion of Languages Used in Messages', fontsize=16)
plt.xlabel('Proportion', fontsize=12)
plt.ylabel('Language', fontsize=12)
plt.tight_layout()
plt.show()
```


Proportion of Languages Used in Messages
Language
Ukrainian
Russian
English
Belarusian
Serbian
Bulgarian
Chinese
Japanese
Kazakh
Mongolian
Macedonian
Kyrgyz
Korean
Italian
Czech
German
Finnish
Nepali
Irish
Romanian
Kinyarwanda
Portuguese
Slovak
Javanese
Spanish
Turkish
Latin
Urdu
Estonian
0.0
0.1
0.2
0.3
0.4
0.5
Proportion