# MATNLARNI HISSIY TAHLIL QILISHDA KORPUS VA LUG‘ATGA ASOSLANGAN YONDASHUVLAR

**Saboxat Allanazarova**[1],
[1]*ToshDO‘TAU tayanch doktoranti*

**KALIT SO‘ZLAR**

fikrlarni aniqlash, hissiyotlar, avtomatik tahlil, statistik modellar, semantik modellar, korpus.

**ANOTATSIYA**

Sentiment (hissiy) tahlil – bu yozma tildan odamlarning fikrlari, baholari, munosabati va hissiyotlarini tahlil qiladigan NLP (tabiiy tilni qayta ishlash) sohasidir. Sentiment tahlil qilish kompyuter lingvistikasi uchun samarali tadqiqot yo‘nalishlaridan hisoblanadi. Avtomatlashtirilgan his-tuyg‘ularni tahlil qilish usullari statistik modellar asosida hissiyotlarni tasniflaydigan ML-algoritmlarini o‘z ichiga oladi. Ushbu maqolada matnlarning avtomatik hissiy tahlilida qo‘llaniladigan ikkita: korpus va lug‘atga asoslangan yondashuvlarning farqi, ustunlik va kamchilik jihatlari haqida so‘z boradi.

**КЛЮЧЕВЫЕ СЛОВА**

анализ тональности, эмоции, автоматизированный анализ, статистические модели, семантические модели, корпус.

**АННОТАЦИЯ**

Сентимент-анализ – это область NLP (обработка естественного языка), которая анализирует мысли, оценки, отношения и эмоции людей на основе письменной речи. Сентимент-анализ – одно из эффективных направлений исследований компьютерной лингвистики. Методы автоматического анализа эмоций включают в себя ML-алгоритмы, которые классифицируют эмоции на основе статистических моделей. В данной статье раскрываются различия, преимущества и недостатки двух подходов к автоматическому сентимент-анализу текстов: корпусного и словарного.

**KEY WORDS**

opinion mining, sentiments, automated analysis, statistical models, semantic models, corpus.

**ABSTRACT**

Sentiment analysis is a field of NLP (Natural language processing) that analyzes people’s thoughts, assessments, attitudes, and emotions based on written speech. Sentiment analysis is one of the most effective areas of computational linguistics research. The methods of automatic emotion analysis include ML algorithms that classify emotions based on statistical models. This article reveals the differences, advantages and disadvantages of two approaches to automatic sentimentalization of texts: corpus and dictionary.

**Kirish.** Korpusga va lug‘atga asoslangan yondashuvlar hissiyotlarni tahlil qilishda qo‘llanadigan ikki xil usuldir. Korpusga asoslangan yondashuv his-tuyg‘ularni ifodalash uchun ishlatiladigan tildagi belgi va tendensiyalarni aniqlash uchun korpus deb nomlanuvchi katta matn to‘plamini tahlil qilishni o‘z ichiga oladi. Ushbu usul ko‘pincha belgi va kontekstga asoslangan matndagi hissiyotlarni avtomatik aniqlash uchun mashinaviy o‘rganish algoritmlaridan foydalanadi. Aksincha, lug‘atga asoslangan yondashuv esa muayyan his-tuyg‘u qutblari (ijobiy, salbiy, betaraf baho) bilan bog‘liq bo‘lgan so‘zlar va iboralarning oldindan belgilangan ro‘yxatiga tayanadi. Ushbu his-tuyg‘u leksikonlari yoki lug‘atlari so‘zlarni va ularga mos keladigan hissiyot ballarini o‘z ichiga oladi va matn qismining hissiyoti undagi so‘zlarning hissiyot ballarini hisoblash orqali aniqlanadi. Boshqacha aytganda, korpusga asoslangan yondashuv his-tuyg‘ularni aniqlash uchun matnning katta to‘plamidagi belgilar va kontekstni tahlil qiladi, lug‘atga asoslangan yondashuv esa matn qismining hissiyotini baholash uchun oldindan belgilangan so‘zlar ro‘yxati va ularning hissiyot ballaridan foydalanadi. Ikkala yondashuv ham o‘ziga xos kamchilik va ustunlikka ega bo‘lib, hissiyotlarni tahlil qilishning aniqligini oshirish uchun birgalikda ishlatilishi mumkin.

**Tadqiqot metodologiyasi.** Tadqiqotni amalga oshirishda tilshunoslikning umumiy va xususiy metodlaridan foydalanilgan. Xususan, ishda LIWC, tematik modellashtirish, PyMorphy kabi kompyuter tilshunosligi metodlari va kvantitativ, korpus ma’lumotlarini statistik tahlil qilish kabi metodlar qo‘llangan.

**Natijalar.** Korpusga asoslangan his-tuyg‘ularni tahlil qilish sizga biznesda mijozlarga xizmat ko‘rsatishni bir necha usul bilan yaxshilashga yordam beradi. Birinchidan, u ijtimoiy media, elektron pochta, chat yoki so‘rovlar kabi turli kanallar bo‘ylab mijozlaringizning

umumiy kayfiyatini kuzatish va o‘lchashga yordam beradi. Bir tomondan, “korpusga asoslangan” yondashuv korpusni misollar manbayi sifatida ko‘radi. Ushbu yondashuvda leksikograf so‘z nimani anglatishi va qanday ishlatilishini sezgisiga tayanib, so‘ngra har bir ma’no va qo‘llanish misollarini topish uchun korpusga murojaat qilishi mumkin.

Lug‘atga asoslangan hissiyotlarni tahlil qilish – bu so‘zlarning sentimental qutblari orqali matn yoki hujjatlarning sentimental holatlari haqida izoh berishga imkon beruvchi matnni qayta ishlash dasturi. So‘nggi yillarda u marketing, sog‘liqni saqlash, ta’lim, turli maqsadlarda foydalanish sababli turli sohalarda mashhur bo‘ldi. Ushbu yondashuvda dastlab bir nechta so‘zlarni olib, lug‘at tuziladi. Keyin onlayn lug‘at, tezaurus yoki WordNet ushbu so‘zlarning sinonimlari va antonimlarini o‘z ichiga olgan holda lug‘atni kengaytirish uchun ishlatilishi mumkin. Jarayon ushbu lug‘atga yangi so‘zlar qo‘shilmaguncha kengaytiriladi. Lug‘atga asoslangan matnning miqdoriy tahlili. So‘z chastotasi va tf-idf matnga asoslangan ma’lumotlarni tekshirishning informatsion usuli bo‘lishi mumkin bo‘lsa-da, boshqa juda mashhur usullar tadqiqotchiga ma’lum ma’no yoki qiymat berilgan har bir hujjatda paydo bo‘ladigan so‘zlar sonini hisoblashni o‘z ichiga oladi.

Lug‘atga asoslangan tokenizatsiya NLPda matnni oldindan belgilangan lug‘at asosida tokenlarga bo‘lish uchun qo‘llaniladigan keng tarqalgan usuldir. Lug‘atga asoslangan tokenizatsiya – bu tabiiy tilni qayta ishlash (NLP) texnikasi bo‘lib, matnni ko‘p so‘zli iboralarning oldindan belgilangan lug‘ati asosida alohida tokenlarga bo‘lish kiradi. Sintaktik va semantik tahlil bu – tabiiy tilni qayta ishlashda qo‘llaniladigan ikkita asosiy usuldir. Sintaktik tahlil so‘zlarning grammatik ma’noga ega bo‘lishi uchun gapdagi joylashuvi. NLP grammatik qoidalarga asoslangan tildan ma’noni baholash uchun sintaksisdan foydalanadi. NLP uchun ishlatilishi mumkin bo‘lgan algoritmlar quyidagilardir: mashinaviy o‘qitish, Vektorli mashinalarni qo‘llab-quvvatlash (SVM), Bayes tarmoqlari, Maksimal entropiya, neyron tarmoq (NN). Naive Bayes Classifier, SVM (Support Vector Machine), Logistik Regressiya, Random Forest va GBM (Gradient Boosting Machines) kabi statistik mashinaviy o‘qitish modellari hissiyotlarni tahlil qilish uchun ahamiyatlidir, ularning har biri o‘zining kuchli tomonlariga ega.

**Statistik yondashuv.** Agar so‘z vaqti-vaqti bilan ijobiy matnlar orasida paydo bo‘lsa, unda uning qutbliligi ijobiydir. Agar so‘z salbiy matnlar orasida tez-tez uchrasa, uning qutbliligini salbiy deb hisoblash mumkin. Agar u ekvivalent hodisaga ega bo‘lsa, uni neytral so‘z deb hisoblash mumkin [5]. Shunday qilib, agar ikkita so‘z bir xil kontekstda ko‘pincha birga kelsa, ular bir xil qutbga ega bo‘lish ehtimoli katta. Shuning uchun, noma’lum so‘zning qutbliligini boshqa so‘z bilan birgalikda paydo bo‘lishning o‘rtacha chastotasini hisoblash orqali tartibga solish mumkin. Buni [8]da ko‘rsatilgan misolda PMI (Pointwise Mutual Information) yordamida amalga oshirish mumkin. Nutqning bir qismidan foydalangan holda, bu usul bigrammalarni ajratib olish orqali matnni tasniflaydi. Keyin PMI har bir bigram uchun qutblilik ko‘rsatkichidan foydalangan holda hisoblanadi.

**Semantik yondashuv.** Ushbu yondashuv semantik jihatdan yaqin so‘zlarga mos keladigan hissiyot qiymatlarini belgilaydi[3]. Semantik jihatdan yaqin so‘zlarga his-tuyg‘u so‘zlari ro‘yxatini olish va sinonimlar va antonimlar bilan boshlang‘ich to‘plamni iterativ ravishda kengaytirish va keyin noma’lum so‘zning hisssiy qutbliligini ushbu so‘zning ijobiy va salbiy

sinonimlarining nisbiy soniga qarab aniqlash mumkin Mijozlarning fikr-mulohazalarini semantik jihatdan tasniflash uchun domendan mustaqil qoidaga asoslangan usul asosida amalga oshiriladi. Bu na o‘rganishga, na leksikaga asoslangan yondashuv bo‘lgani uchun, bu usulni boshqalar bilan solishtirish qiziq. Ushbu usul uch qismda ishlaydi. Birinchidan, sharhlar oldindan qayta ishlanadi: ular tuzatiladigan jumlalarga bo‘linadi va jumlaning har bir so‘zini belgilash va saqlash uchun “Nutqning bir qismi” (POS) usuli qo‘llaniladi [4]. Ikkinchidan, kontekstual ma’lumotlarga va jumla tuzilishiga asoslanib, gapning qutbliligini aniqlashga imkon beruvchi fikr so‘zini ajratib olish bosqichi. Mahsulotning “aspektlari” jumlalarning ot iboralari sifatida aniqlanadi. Oxirgi qism qoidaga asoslangan modul yordamida jumlalarni obyektiv yoki subyektivga tasniflashdan iborat. Har bir fikr so‘zi SentiWordNet lug‘atidan 117 662 dan ortiq so‘zdan iborat semantik ballni o‘z ichiga olgan semantik ballga ega. Ushbu ballar bilan har bir muddatni baholash orqali, sharh ijobiy yoki salbiy ekanligini aniqlash uchun hukmga jumla darajasida ball berilishi mumkin.

**So‘zning bog‘liqligi.** Ikki so‘zning ma’nosi o‘xshashlikdan boshqa yo‘llar bilan ham bog‘lanishi mumkin [2]. Bunday bog‘lanish sinflaridan biri so‘z birikmasi. Kofe va chashka so‘zlarining ma’nolarini ko‘rib chiqamiz. Qahva stakanga o‘xshamaydi; ular deyarli hech qanday xususiyatga ega emas (qahva o‘simlik yoki ichimlik; stakan esa ma’lum bir shaklga ega bo‘lgan, ishlab chiqarilgan obyektdir). Lekin qahva va chashka aniq bog‘liq; ular kundalik hayotda (chashkadan qahva ichish uchun foydalaniladi) birgalikda ishtirok etish bilan bog‘liq. Xuddi shunday, skalpel va jarroh o‘xshash emas, lekin bir-biriga bog‘liqdir (jarroh skalpeldan foydalanadi).

Agar so‘zlar o‘rtasida umumiy bog‘liqlik bo‘lsa, ular bir semantik sohaga tegishli bo‘ladi. Semantik maydon – bu ma’lum bir semantik sohani qamrab oluvchi va bir-biri bilan tuzilgan munosabatlarga ega bo‘lgan so‘zlar to‘plami. Masalan, shifoxona (jarroh, skalpel, hamshira, dori, kasalxona), restoran (ofitsiant, menyu, idish, ovqat, oshpaz) yoki uy (eshik, tom, oshxona, to‘shak) kabi so‘zlar o‘z atrofiga bir nechta so‘zlarni biriktirib, semantik maydon hosil qilishi mumkin. Semantik maydonlar, shuningdek, matndagi bog‘langan so‘zlar to‘plamini induktsiya qilish uchun katta matnlar to‘plamida nazoratsiz o‘rganish (Unsupervised machine learning) aniqlash uchun juda foydali vositadir. Bunday bog‘liqlikni bilish matnni tahlil qilish hamda mashina tarjimasida foydali bo‘lishi mumkin. Nihoyat so‘zlar nominativ va konnatativ ma’noga ega bo‘lib, konnotativ so‘z turli gaplarda turli xil ma’nolarga ega, ammo bu yerda so‘z ma’nosining yozuvchi yoki o‘quvchining his-tuyg‘ulari, hissiyotlari, fikrlari yoki baholashlari bilan bog‘liq tomonlariga e’tibor qaratamiz. Masalan, ba’zi so‘zlar ijobiy ma’noga (baxtli, begunoh), boshqalari esa salbiy ma’noga ega (qayg‘uli, qalbaki). Ijobiy yoki salbiy bahoni ifodalovchi so‘zlar va birikmalar hissiy (sentiment) deb ataluvchi tahlilning predmeti bo‘lib, u NLPning sohasi hisoblanadi[1].

**1-jadval. Leksemaga asoslangan yondashuvning yutuq va kamchiliklari**

| Afzalliklari: | Kamchiliklari |
|---|---|
| Ko‘plab ochiq manbalar (masalan, SentiWordNet, SenticNet, WordNet) mavjud | Ijtimoiy tarmoqdagi ma’lumotlarni tasniflashda murakkablik |
| Tejamkor, chunki ular hissiyotlarni tahlil qilish algoritmlarini talab qilmaydi | Kinoya va istehzoni farqlamaydi |
| O‘quv ma’lumotlariga ehtiyoj qolmaydi, chunki so‘zlarning ma’nosiga tezkor kirish imkoniyati mavjud | Grammatik xatolar, noto‘g‘ri imloni aniqlamaydi |
| | Juda qat’iy va domenga bog‘liq (so‘z kontekstdan kelib chiqib baholanmaydi) |

Hisoblash tilshunosligida ma’noga asosiy yondashuv til ma’lumotlarining ma’no bilan bog‘liq mazmunini qamrab oluvchi rasmiy ma’no ko‘rinishlarini ishlab chiqishni o‘z ichiga oladi. Ushbu taqdimotlar tildan dunyoni umumiy ma’noda bilishgacha bo‘lgan bo‘shliqni to‘ldirishga mo‘ljallangan. Bu tasvirlarning sintaksisi va semantikasini aniqlovchi ramkalar ma’no ifodalash tillari deb ataladi. Bunday tillarning xilma-xilligi tabiiy tillarni qayta ishlash va sun’iy intellektda qo‘llaniladi [7].

**Muhokama va xulosa.** Hozirgi kunda biz hayotimizni ijtimoiy tarmoqlarsiz tasavvur qila olmaymiz. Har kim internetdan turli maqsadlarda, ya’ni ma’lumot qidirish yoki biror narsa joylashtirish uchun foydalanadi. Fikr ma’lumotlarining keng mavjudligi fikrlarni izlash va tasniflashning avtomatlashtirilgan tizimini yaratish zaruratini keltirib chiqardi. Sentiment tahlili katta hajmli ma’lumotlarni avtomatik tasniflash va baholash uchun ishlatiladi.

**Foydalanilgan adabiyotlar ro‘yxati:**

1. Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets / Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2017. – PP. 2–12.

2. Devika M. D., Sunitha C., Ganesh A. Sentiment Analysis: A Comparative Study On Different Approaches // Procedia Computer Science, 2016. – Vol. 87. – PP. 44–49.

3. Serrana-Guerrero J., Olivas J. A., Romero F. P., E. Herrera-Viedma. Sentiment analysis: A review and comparative analysis of web services // Information Sciences, 2015. – Vol. 311. – PP. 18-38.

4. Lui B. Sentiment Analysis and Subjectivity / Handbook of Natural Language Processing, N. Indurkhya and F. J. Damerau, Eds. 2nd ed., 2010.

5. Matlatipov S., Kuryozov E., Miguel A. A., Corlos-Rodriguez. Deep learning vs. classic models on a new Uzbek sentiment analysis dataset / Proceedings of the 9th Language & Technology conference: Human language technologies as a challenge for computer science and linguistics, Poznan, 2019. – PP. 258-262.

6. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A

survey // Ain Shams Engineering Journal, 2014. – Vol. 5. – № 4. – PP. 1093-1113.

7. Baccianella S., Esuli A., Sebastiani F. SEN TI WORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining / Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010.

8. Neviarouskaya A., Prendinger H., Ishizuka M. Recognition of Affect, Judgment, and Appreciation in Text / Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, 2010. – PP. 806–814.