VILNIAUS UNIVERSITETAS
Fiziniai mokslai, Informatika (09P)

INTELEKTINIŲ METODŲ TYRIMAI NUOMONIŲ ANALIZEI DIDELIUOSE DUOMENŲ MASYVUOSE

III ataskaitiniai metai

2015m. spalio mėn. 1d. – 2019m. rugsėjo mėn. 30d.

Dokt.: Konstantinas Korovkinas
Darbo vadovas: prof. dr. Gintautas Garšva

Pavadinimas

Patikslintas disertacijos pavadinimas:

“Hybrid K-means and SVM method for sentiment polarity classication in Large Scale text arrays”

“Hibridinis K-means ir SVM metodas nuomonių poliariškumo klasifikavimui didelės apimties tekstiniuose masyvuose”

Tyrimo objektas, tikslai, planuojami gauti rezultatai

Tyrimo objektas:

Intelektiniai metodai.

Tyrimo tikslas:

Sukurti hibridinį intelektinį metodą nuomonių analizei dideliuose duomenų masyvuose.

Tyrimo uždaviniai:

- Apžvelgti esamus nuomonių/sentimentų analizės tyrimus ir identifikuoti problemą.
- Sukurti hibridinį intelektinį metodą ir pagrįsti naujas savybes, požymius.
- Sudaryti programinį prototipo projektą, adaptuojant sukurtą metodą.

Planuojami rezultatai:

Programinis prototipo projektas nuomonių analizei dideliuose duomenų masyvuose, adaptuojant sukurtą metodą.

Ataskaitinių metų darbo planas

Studijų planas:

Mokslinių tyrimų planas:

- Intelektinių metodų tyrimas nuomonių analizei dideliuose duomenų masyvuose, adaptuojant sukurtą metodą.
- Programinis prototipo projektas, tyrimai, testavimas, galimi praktinės realizacijos variantai.

Rezultatų pristatymo planas:

- Dalyvavimas Kauno fakulteto doktorantų tarpdisciplininiame seminare.
- Dalyvavimas konferencijoje “Information Society and University Studies” (IVUS) 2018.
- Dalyvavimas tarptautinėje mokslinėje konferencijoje.

Mokslinių publikacijų planas:

Planuojamas mokslinis straipsnis „SVM and Naïve Bayes Classification Ensemble Method for Sentiment Analysis“.

Ataskaita

2015–2017 m. m. išlaikyti egzaminai:

	Dalyko pavadinimas	Kreditų skaičius ECTC	Atsiskaitymo data	Dalyko konsultantas	Įvertinimas
1	Informatikos ir informatikos inžinerijos tyrimo metodai ir metodika	9	2016.06 (2016.06.09)	Prof. dr. A. Čaplinskas	**8**
2	Skaitinis intelektas	7	2016.11 (2016.09.12)	Doc. dr. V. Rudžionis	**9**
3	Sistemų analizės technologijos	7	2017.03 (2016.10.20)	Prof. dr. S. Gudas	**9**
4	Duomenų analizės strategijos ir sprendimų priėmimas	7	2017.06 (2016.12.19)	Prof. habil. dr. G. Dzemyda dr. O. Kurasova dr. J.Bernatavičienė	**9**

Ataskaita

2015–2018 m. m. disertacijos rengimo planas:

	Etapo pavadinimas	Planuota atlikti data	Pastabos
1	Mokslinių tyrimų disertacijos tema apžvalga ir analizė: 1. Intelektinių metodų dideliuose duomenų masyvuose tyrimų apžvalga. 2. Nuomonių/sentimentų analizės tyrimų apžvalga. 3. Problemų identifikavimas. 4. Mokslinė hipotezė.	2016.09	Įvykdyta. Studijų eigoje bus pildoma.
2	2.1. Teorinis tyrimas: 1. Siūlomas intelektinis metodas ar jų hibridas, naujų savybių, požymių pagrindimas. 2. Kuriamo metodo palyginimas su jau egzistuojančiais metodais.	2016.12	Įvykdyta. Studijų eigoje bus pildoma.
3	2.2. Tyrimo metodikos sudarymas: 1. Tyrimo metodikos parinkimas. 2. Teorinio ir eksperimentinio tyrimų planavimas, sukurtąjam intelektiniam metodui.	2017.09	Įvykdyta. Studijų eigoje bus pildoma.

Ataskaita

2015–2018 m. m. disertacijos rengimo planas:

	Etapo pavadinimas	Planuota atlikti data	Pastabos
4	Eksperimentinis tyrimas: 1. Intelektinių metodų tyrimas nuomonių analizei dideliuose duomenų masyvuose, adaptuojant sukurtą metodą. 2. Programinis prototipo projektas, tyrimai, testavimas, galimi praktinės realizacijos variantai.	2018.09	Įvykdyta. Studijų eigoje bus pildoma.
5	Gautų duomenų analizė, apibendrinimas, išvadų parengimas.	2018.10	Vykdoma. Studijų eigoje bus pildoma.

Ataskaita

Publikacijos 2015–2018 m. m.:

STRAIPSNIAI leidiniuose, įtrauktuose į Mokslinės informacijos instituto (ISI) duomenų bazes

1) Korovkinas, K., Danėnas, P., Garšva, G. *SVM and Naïve Bayes Classification Ensemble Method for Sentiment Analysis*. Baltic Journal of Modern Computing, 5(4), pp.398-409, 2017.

2) Korovkinas, K., Danėnas, P., Garšva, G. *SVM Accuracy and Training Speed Trade-Off in Sentiment Analysis Tasks*. In International Conference on Information and Software Technologies (pp. 227-239). Springer, Cham, 2018.

STRAIPSNIAI recenzuojamuose periodiniuose leidiniuose

3) Korovkinas, K., Garšva, G., *Selection of intelligent algorithms for sentiment classification method creation*. Proceedings of the International Conference on Information Technologies, Vol-2145, Kaunas, Lithuania, pp. 152-157, ISSN 1613-0073, CEUR, 2018.

4) Vaitonis, M., Masteika, S., Korovkinas, K. *Algorithmic trading and machine learning based on GPU*. Proceedings of the Symposium for Young Scientists in Technology, Engineering and Mathematics, Vol-2147, Gliwice, Poland, pp. 49-54, ISSN 1613-0073, CEUR, 2018.

Ataskaita

Dalyvavimas konferencijose 2015–2018 m. m.:

1) International Conference on Information Technologies (IT2018), Kaunas, Lithuania
2) Symposium for Young Scientists in Technology, Engineering and Mathematics (SYSTEM2018), Gliwice, Poland
3) International Conference on Information and Software Technologies (ICIST2018), Vilnius, Lithuania

Dalyvavimas seminaruose 2017–2018 m. m.:

4) Kauno fakulteto doktorantų tarpdisciplininis seminaras Palangoje. Pristatyta disertacija.
5) VU Duomenų Mokslo ir Skaitmeninių Technologijų Instituto organizuotas Informatikos inžinerijos problemų seminaras: "**Mašininis mokymasis ir nuotaikų analizė**". Pristatytas straipsnis "*Atraminių vektorių ir naivaus Bajeso klasifikavimo hibridinio metodo taikymas sentimentų analizei*".

Ataskaita

Kita veikla 2015–2018 m. m.:

2016m. Bakalauro kursinio darbo vadovas (darbas įvertintas 8) ir baigiamojo darbo vadovas (darbas įvertintas 10).

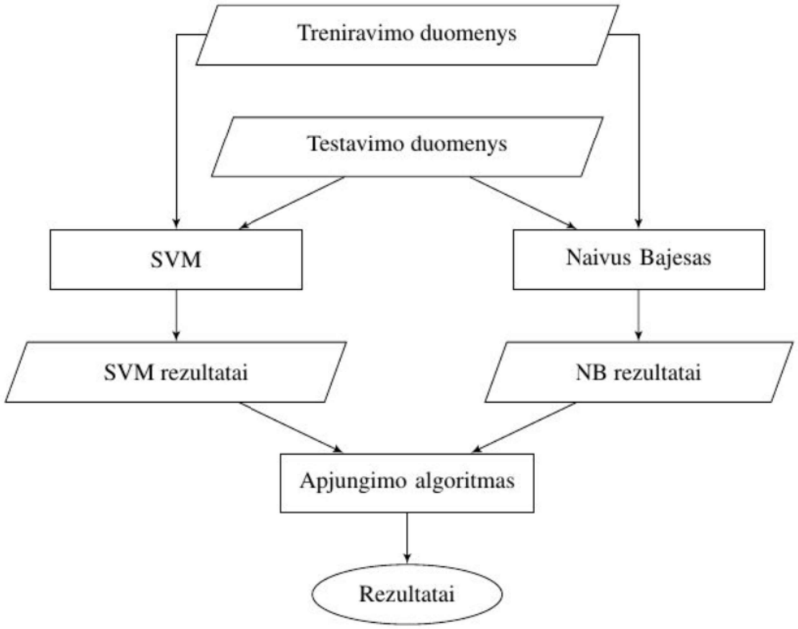
2017m. Bakalauro kursinio darbo vadovas (darbas įvertintas 7) ir baigiamojo darbo vadovas (darbas įvertintas 9).

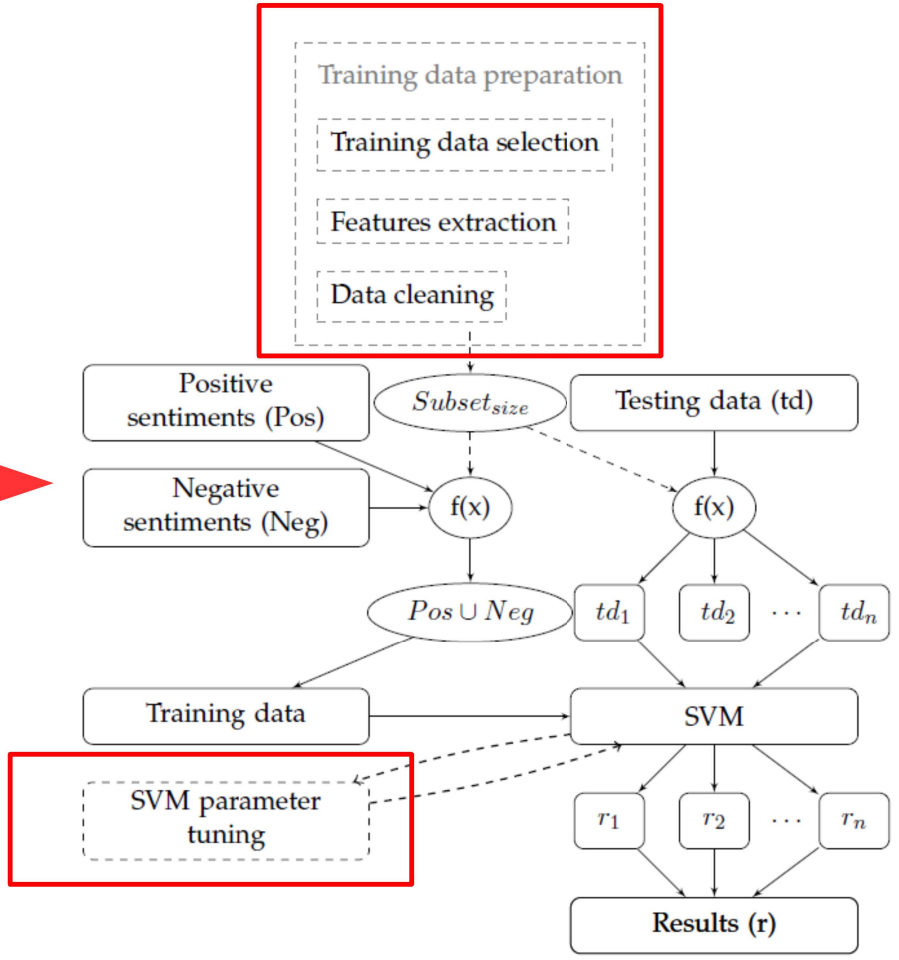
2018m. Kauno fakultete dėstau „Kompiuterių architektūrą“ anglų ir lietuvių grupėms.

Ataskaita

Siūlomas intelektinis metodas:

Rezultatų apjungimo metodas





SVM spartinimo metodas

Ataskaita

Apjungimo algoritmas:

Įvestis: Tarkim th_2 – $R_{SVM}\{p\}$ pasirinkimo slenkstis (algoritmo žingsnis (2)) ir th_3 – $R_{SVM}\{p\}$ pasirinkimo slenkstis (algoritmo žingsnis (3)).

$R_{SVM} = \{SVMsent, v\}$ – SVM rezultatų aibė, *SVMsent* – nuomonė;
$R_{NB} = \{NBsent, v\}$ – Naivaus Bajeso rezultatų aibė, *NBsent* – nuomonė;
p – sakinių klasifikavimo tikimybė;
v – Naivaus Bajeso rezultatų reikšmė: "1", jei "pozityvus" sakinys ir "-1", jei "negatyvus" sakinys;
$th_3 = \min(R_{SVM}\{p\}) + (\sigma_{R_{SVM}\{p\}})/2 - 0.01$ (naudojama mūsų pasiūlyta formulė), kur $\sigma_{R_{SVM}\{p\}}$ yra $R_{SVM}\{p\}$ standartinis nuokrypis;
R – rezultatų aibė.

Rezultatų apjungimas:

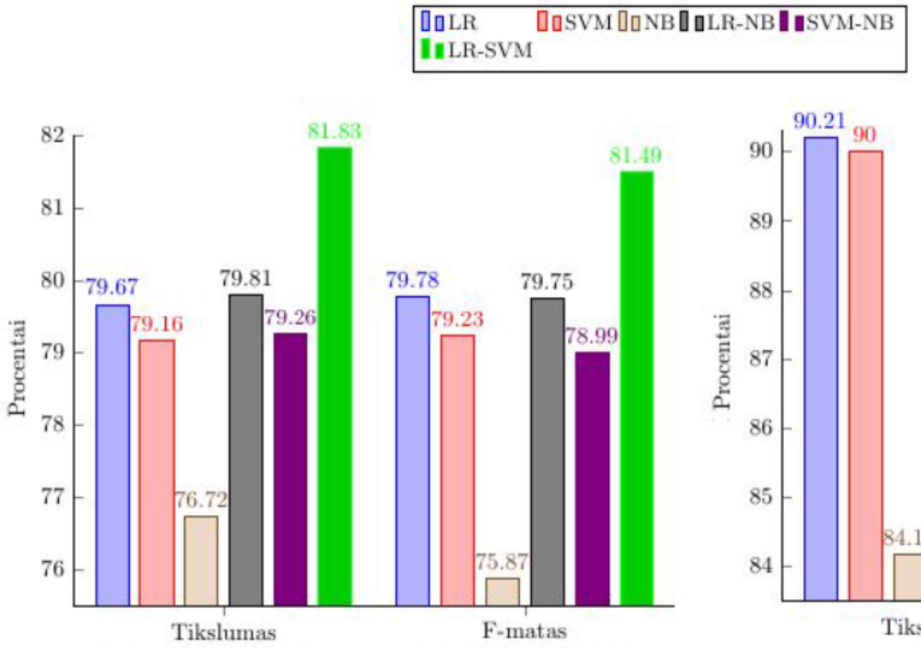
1. Randami vienodi SVM ir Naivaus Bajeso rezultai.
   $R = R_{SVM} \cap R_{NB} = \{x : x \in R_{SVM}\{SVMsent\} \text{and } x \in R_{NB}\{NBsent\}\}$
2. Randami skirtingi SVM ir Naivaus Bajeso rezultai.
   $R_{SVM}\{SVMsent\} \Delta R_{NB}\{NBsent\}$ and $R_{SVM}\{p\} < th_2$
3. $R = \begin{cases} R \cup R_{SVM}, \text{if } |R_{SVM}\{p\}| < th_3 \\ R \cup R_{NB}, \text{if } |R_{SVM}\{p\}| \geq th_3 \end{cases}$

Išvestis: klasifikavimo rezultatų aibė R = {S, nuomonė} ir Tikslumas.

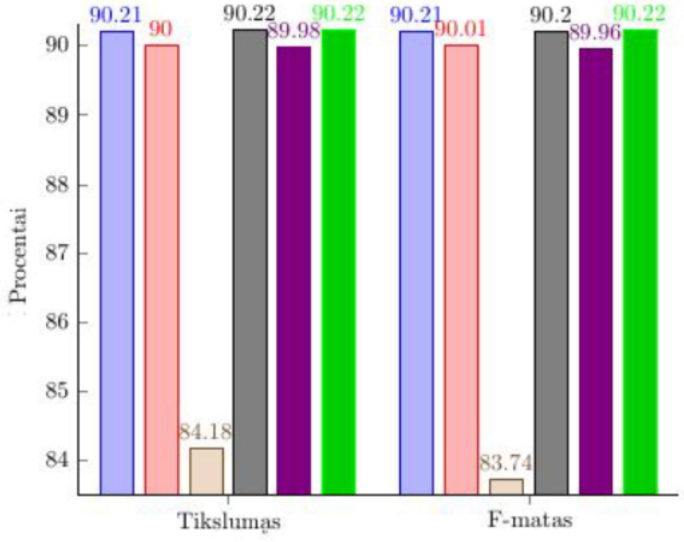
Ataskaita

Rezultatai (rezultatų apjungimo metodas):

Exp. No.	Dataset	Training data 70%	Testing data 30%
1	sentiment140	1.12M	480K
2	Amazon reviews	2.8M	1.2M



Eksperimento rezultatai, kai naudojama Twitter duomenų aibė



Eksperimento rezultatai, kai naudojama Amazon duomenų aibė

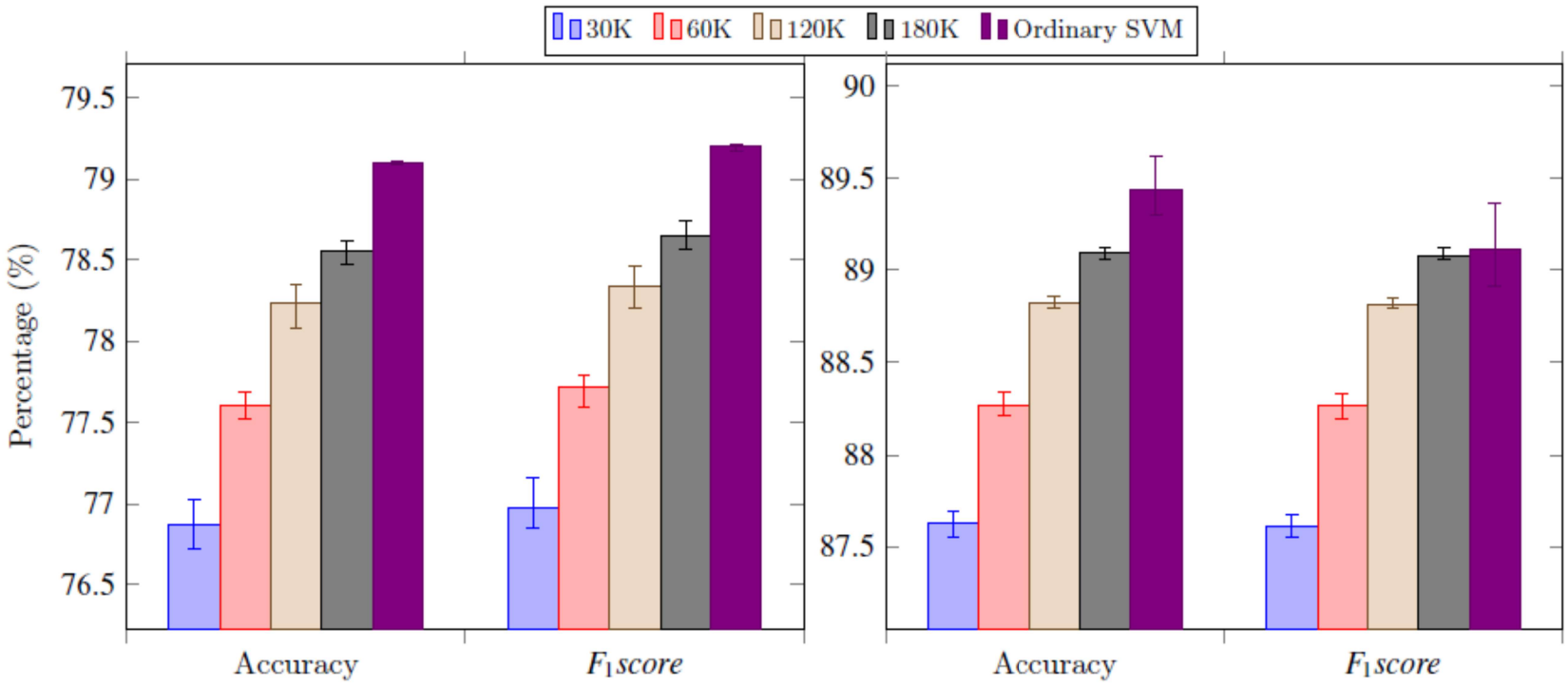
Ataskaita

Rezultatai (SVM spartinimo metodas):

Exp. No.	Dataset	Testing data size (TDs)	Subset size (SubS)	Subsets quantity (SQ) trunc(TDs/Ss)	Remainder TDs-(SubS*SQ)	Calculated training data dependently on SubS
3	sentiment	480K	30K	16	0	70K
4	140	480K	60K	8	0	140K
5		480K	120K	4	0	280K
6		480K	180K	2	120K	420K
7	Amazon	1.2M	30K	40	0	70K
8	reviews	1.2M	60K	20	0	140K
9		1.2M	120K	10	0	280K
10		1.2M	180K	6	120K	420K

Ataskaita

Rezultatai (SVM spartinimo metodas):

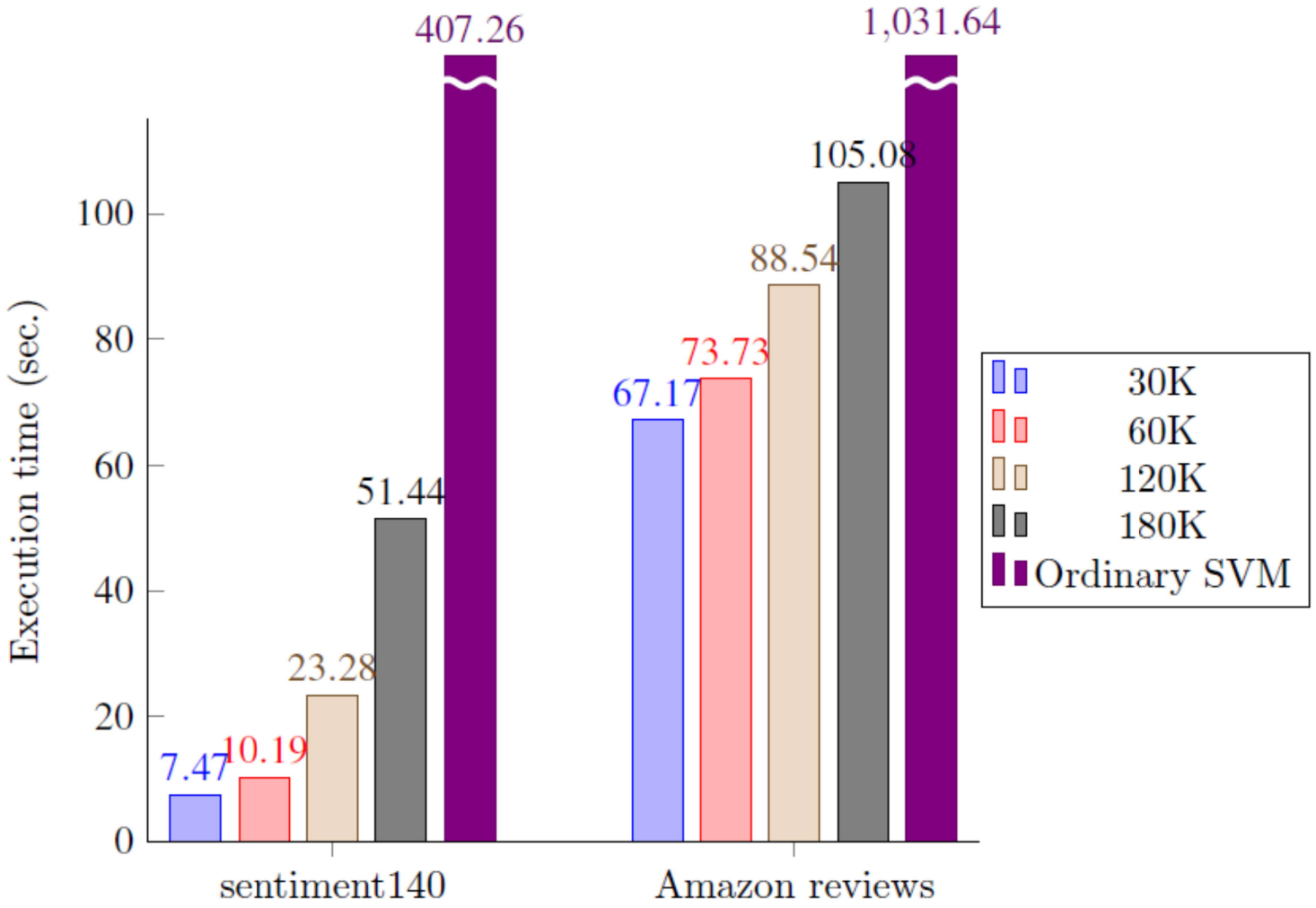


sentiment140 results

Amazon reviews results

Ataskaita

Rezultatai (SVM spartinimo metodas):



Išvados

- Eksperimentai parodė, kad pasiūlytas rezultatų apjungimo metodas parodė geriausius rezultatus, kai buvo naudojami beveik lygūs klasifikatoriai (mūsų atveju Logistinė regresija ir Atraminių vektorių modeliai). LR-SVM (ACC) 81,83% ir 90,22%, kai naudojamas vien LR 79,67% ir 90,21%. Stipriausio ir silpniausio metodo (Logistinė regresija ir naivus Bajesas) rezultatai taip pat yra geresni už pavienio LR metodo rezultatus, nors ir nežymiai. LR-NB (ACC) 79,81% ir 90,22%.
- Atraminių vektorių mašinų spartinimo metodas tikslumu nežymiai nusileido paprastam atraminių vektorių metodui, bet vykdymo laikas, kai buvo naudojama the Stanford Twitter sentiment corpus duomenų aibę buvo 7.9-54x didesnis, o Amazon customer reviews duomenų aibės atveju 9.8-15.35x didesnis.

Darbo planas kitai atestacijai

Mokslinių tyrimų planas:

	Etapo pavadinimas	Atlikimo terminas
1	Gautų duomenų analizė, apibendrinimas, išvadų parengimas	2018.10
2	Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas.	2019.04

Papildyti ir pataisyti pristatytas disertacijos dalis.

Darbo planas kitai atestacijai

Mokslinių publikacijų planas:

- Planuojamas mokslinis straipsnis "*SVM and k-Means Hybrid Method for Large Scale Sentiment Analysis*".
- Planuojamas mokslinis straipsnis.

Dalyvavimas konferencijose, seminaruose, kitose doktorantų mobilumo veiklose:

- Dalyvavimas tarptautinėje konferencijoje „Duomenų analizės metodai programų sistemoms" (DAMSS).
- Dalyvavimas VU Duomenų Mokslo ir Skaitmeninių Technologijų Instituto organizuojamame seminare. Planuojamas pirminio disertacijos varianto pristatymas.

Ačiū už dėmesį