АВТОМАТИЧНО РАЗПОЗНАВАНЕ НА ЕМОЦИИ В ТЕКСТ

Тодор Новаков, Иван Койчев

Факултет по Математика и Информатика,
Софийски Университет „Св. Климент Охридски"
todor.novakov3@gmail.com, koychev@fmi.uni-sofia.bg

***Резюме**: С бързото развитие на информационните технологии и разширяването на Интернет, потребителите на социалните мрежи и блоговете генерират огромно количество текст, наситен с емоции. Автоматичното разпознаване на емоциите в текста е от интерес както на социалните науки, така и за приложенията за откриване на отношението към даден продукт. Тази статия представя и изследва методите за машинното обучение, използвани за разпознаване на емоциите в текста. Направено е сравнително изследване на ефективността им за решаване на тази задача. Представени са и резултатите от измерване на връзките (корелациите) между основните типове емоции в текста.*

***Ключови думи**: машинно самообучение, многоетикетна класификация, класификация на емоции*

1. Увод

Анализът на мнения и емоции (Opinion Mining and Sentiment Analysis) са научни области, чиято цел е да извлече знание за отношението на субект към тема, продукт, събитие или др. С помощта на алгоритми за машинно самообучение (Machine Learning), подходи за обработка на естествен език (Natural Language Processing), откриване на знания в текст (Text Mining), статистически методи и др. се открива, анализира и оценява емоционалната наситеност в обекта на изследването - текст, видео или звук. Съществува информация за изследвания в тази област още от началото на текущия век. Настоящата разработка се фокусира върху класификация на емоционално зареден текст в гама от 11 основни емоции.

Основните видове класификации при анализирането на текст са: двоична (binary classification), многокласова (multi-class classification) и многоетикетна (multi-label classification). При двоичната класификация, даден текст се определя само като положителен или отрицателен. При многокласовата класификация текстът се класифицира в точно един клас измежду много класове (повече от два). Пример за такава класификация е когато текст се определя като положителен, отрицателен или неутрален. За разликата от предходните две, при многоетикетната класификация – текстът се класифицира едновременно в повече от един клас т.е. в него се откриват емоции като оптимизъм, наслада, щастие и др.

Многоетикетна класификация се характеризира като обобщение на многокласовата класификация - всеки пример може да притежава повече от един етикет. Два са основните метода за решаването на такава задача-трансформация на задачата до няколко задачи от другите два вида и използване на адаптирани алгоритми.

Трансформирането на многоетикетна задача се свежда до решаването на известен брой едноетикетни задачи т.е. двоични класификационни задачи или няколко многокласови. Най-често използваните методи са двоична проверка (Binary Relevance), верига от класификатори (Classifier Chains) и метода на всички подмножества от етикети (Label Powerset). За разлика от предходния подход - адаптираните алгоритми решават многоетикетна задача без да я разделят на множество от подзадачи. Примери за често използвани алгоритми от този тип са подобренията на метода на най-близкия съсед (k-Nearest Neighbour (kNN)) [1].

2. Предишни изследвания

2.1. Класификация на текст според емоциите

В последното десетилетие анализът на чувства в текст е обект на редица изследвания. Aman and Szpakowicz [2] провеждат изследване върху класификация на блог постове като ги класифицира като емоционални или не. Анализират резултатите от Наивен Бейсов класификатор (Naive Bayes) и метод на опорните вектори (Support Vector Machines (SVM)). Alm et al. [3] разширява задачата до многокласова, като класифицира текст като положителен, отрицателен и неутрален. Подобно проучване осъществява Yang et al. [4] - използват се четири класа - "happy" (щастлив), "joy" (радост), "sad" (тъжен) и "angry" (ядосан). Сравнява се ефективността на метода на опорните вектори със статистическия метод на условни случайни полета (Conditional Random Field (CRF)). S. Mohammad and F. Bravo-Marquez [5] в своя труд за многокласова класификация на емоции освен сравнителния анализ на алгоритмите, те правят такъв и върху избора на атрибути.

Подходът, който се разглежда в настоящият текст, представя както основните методи за решаване на многоетикетна задача, така и сравнителен анализ на използваните алгоритми за класификация.

S. M. Liu, J.Chen [6] фокусират своето изследване върху ефективността на доста голям набор от алгоритми и използването на различни метрики за тяхното оценяване. За разлика от предходните експерименти този се извършва върху многоетикетна класификация на емоции.

Интересен експеримент, но в областта на музиката, се провежда от следните автори Trohidis, Tsoumakas, Kalliris,Vlahavas [7]. При него се разглеждат основните методи и алгоритми за решаване на такъв тип задача.

Крайчев и Койчев [10] предлагат метод за автоматична класификация на българските прилагателни имена в пространство от предварително избрани емоционални оси като: любов – омраза, щедрост – алчност, добрина – злина и др. В основата на изследването стоят данните за честотата на срещане на думите в документи от индекса на машина за търсене bing. Резултатите от изследването отразяват съвременното използване на българския език в глобалната мрежа.

2.2. Подходи за решаване на многоетикетна класификация

Известни са два основни подхода за многоетикетна класификация - трансформация на задачата и използване на адаптирани алгоритми. Трансформацията на многоетикетна класификационна задача се свежда до решаването на няколко многокласови или двукласови задачи. Най-известният метод, който се използва е двоична проверка (Binary Relevance). При него се обучава по един двоичен класификатор за всеки етикет, който определя дали стойността на този етикет е положителна или отрицателна. Методът се характеризира като теоретично прост и интуитивен. Като недостатък на този подход може да се посочи загубата на корелацията между отделните етикети [8]. Съществуват и подобрения на този метод, които запазват зависимостта между отделните класове.

Друг често използван метод е верига от класификатори (Classifier Chains) - комбинира ефикасността на двоична проверка (Binary Relevance), като запазва зависимостта между отделните класове. Обучават се N на брой класификатори, които са свързани в последователност в пространството от атрибути. Като пространството от атрибутите, което се подава на класификатор i+1, е разширено с двоичната стойност, която е пресметната от класификатор i за конкретен етикет [8]. Като недостатък на този метод може да се посочи, че подредбата на етикетите може да доведе до намаляване на точността на предсказване.

Методът на всички подмножества от етикети (Label Powerset) трансформира задачата до многокласова класификация като един класификатор се обучава за всяка уникална комбинация от етикети, извлечени от данните за обучение. Така всяка комбинация се третира като един атомарен клас. Въпреки че методът на всички подмножества от етикети запазва корелацията между отделните класове, той се характеризира с голяма изчислителна сложност в най-лошия случай, както и с прекомерно нагаждане (overfitting) към данните за обучение, защото различните комбинации от етикети се генерират единствено на базата на тях.

Адаптираните алгоритми са друг основен метод за решаване на многоетикетна класификация. Използват се алгоритми, които директно решават тази задача без да я трансформират. Най-често се използват подобрения на метода за най-близкия съсед (kNN): MLkNN [1] и BRkNN [9]. MLkNN разширява kNN като за всеки пример, след определяне на множеството от негови съседи, се избира класа с най-голямата постериорна вероятност (Maximum A Posteriori (MAP)), въз основа на етикетите на неговите съседи.

Алгоритъмът BRkNN комбинира силните страна на kNN и метода двоична проверка (Binary Relevance) с подобрена изчислителна сложност. Независимите предсказвания се осъществяват за всеки етикет от множеството на най-близките съседи [9].

3. Експеримент и резултати

Цел на проведените експерименти е да се изследва приложимостта на алгоритмите за машинно самообучение за учене на многоетикетни класификатори на емоционално зареден текст. Също така цели се и изследване на връзките (корелациите) между различните емоции с идеята, че те могат да доведат до подобряване на точността на класификация.

3.1. Данни

За целта на текущия експеримент са използвани данни от състезанието SemEval-2018 Task, Affect in Tweets, Task E-c: Detecting Emotions. Данните представляват мнения, които са събрани от една от големите социални мрежи в уеб - Twitter. Всяко мнение е анотирано със следните етикети: "anger" (гняв), "anticipation" (очакване), "disgust" (отвращение), "fear" (страх), "joy" (радост), "love" (любов), "optimism" (оптимизъм), "pessimism" (песимизъм), "sadness" (тъга), "surprise" (изненада) и "trust" (доверие). Съществува и още един етикет, който не е определен явно - "neutral or no emotion" (неутрален или неемоционален). Етикети представляват основните осем емоции, дефинирани от R. Plutchik (1980), като са добавени и често срещаните в социалните мрежи: "love" (любов), "optimism" (оптимизъм), and "pessimism" (песимизъм).

3.2. Предварителна обработка на данни

Извършена е предварителна обработка на текста като са премахнати пунктуацията, стоп думите, както и цитираните потребителски имена от Twitter. Също така текстът е изчистен и от малки картинки, които се вмъкват в текст. Хаштаговете представляват думи или последователност от думи, които започват със символа "#". В настоящия текст те са запазени, като единствено е премахнат предхождащия ги символ.

3.3. Избор на атрибути

Ключова роля в машинното самообучение е правилният избор на атрибути за обучение на класификатора. Най-честият подход при обработка на текст е представянето му като множество от думи (Bag of Words). Такъв тип представяне има недостатък - губи се важна семантична и синтактична информация за текста. За да се запази контекста в мненията се прибягва до използването на n-грами, последователности от n на брой символа. В настоящия експеримент се използват различни по дефиниция n-грами.

N-грама ще наричаме последователност от n на брой символа в реда, в който се срещат в текста. Аналогично на предходното, n-думи представлява последователност от n на брой думи. Ще дефинираме и n-грами в думата, които са еквивалентни на n-грамите, но последователността от символите е в границата на дадена дума.

Всяка „n-грама", „n-дума", „n-грама в думата" се разглежда като отделен атрибут. За определяне стойностите на тези атрибути се използва мярката TF-IDF, която е широко използвана в Извличането на информация. Тази оценка представлява числова стойност, която относително определя колко често дадена n-грама/дума/грама в думата се среща в текст.

3.4. Метрики за оценяване

За оценяване на модела са използвани основните метрики в машинното самообучение – прецизност (precision), възвращаемост (recall) и мярката F-measure, която е претеглено хармонично средно на предходните две. При многокласова и многоетикна класификации метриката F1 се разглежда в два варианта - micro-F1 и macro-F1. Micro-F1 оценява класификатора като събира броя на правилно положително класифицираните примери (true positives), грешно положително класифицираните примери (true negatives), правилно отрицателно класифицираните примери (false positives) и грешно отрицателно класифицираните примери (false negatives) за всяка една класификация. Така обобщени метрики се използват за изчисляване на прецизност, възвращаемост и F1 за цялата задача. Macro-F1 събира пресметнатите метрики прецизност и възвращаемост за всеки клас или етикет и намира тяхното средно аритметично. Предпочитана е оценката micro-F1, тъй като взима предвид неравномерното разпределение на класовете или етикетите.

3.5. Оптимизация на параметрите

За целта на настоящия експеримент ще се фокусираме конкретно върху оптимизация на оценката micro-F1. При избора на атрибути се експериментира с различна дължина на n-грамите/думите/грами в думите. При метода за трансформация на задачата се използват различни алгоритми за

класификация, а при адаптираните алгоритми се задават различни стойности за параметър за брой най-близки съседи.

3.6. Трансформация на задачата

При експеримента за трансформация на задачата са използвани следните алгоритми за машинното самообучение от тип учене с учител (supervised learning) - Наивен Бейсов класификатор (Naive Bayes classifier), логистична регресия (Logistic regression) и метод на опорните вектор (SVM). Използван е варианта на полиноминалния Наивния Бейсов класификатор, който е подходящ за дискретно множество от атрибути.

Метод	Кл. алго-ритъм	Бр. думи в n-грама	Precision	Recall	micro-F1
Binary Relevance	*Multinomial NB*	*[1, 4]*	***0**.466*	*0.483*	*0.474*
Classifier Chains	*SVM*	*[1, 1]*	***0.550***	*0.470*	***0.507***
Label Powerset	*SVM*	*[1, 1]*	*0.493*	***0.497***	*0.495*

Таблица 1 Трансформация на задачата. N-грама от думи.

На таблица 1 са представени резултати от експеримента, сравнени при избор на параметъра за n-грама, съответстващ на n-дума. Най-висока прецизност и micro-F1 се постига при метода верига от класификатори (Classifier Chains), а възвращаемост при метода на всички подмножества от етикети (Label Powerset). При по-горе посочените методи параметърът за дължина на n-дума е оптимизиран до дължина 1 т.е. методите работят най-добре с цели думи.

Метод	Кл. алго-ритъм	Бр. символи в n-грама	Precision	Recall	micro-F1
Binary Relevance	*Multinomial NB*	*[2, 7]*	***0.542***	***0.480***	***0.509***
Classifier Chains	*SVM*	*[2, 3]*	*0.537*	*0.451*	*0.490*
Label Powerset	*SVM*	*[3, 5]*	*0.500*	***0.480***	*0.490*

Таблица 2 Трансформация на задачата. N-грама от символи.

На таблица 2 при избора на n-грама от поредица от символи най-добри резултати по всички метрики се постигат от метода двоична проверка (Binary Relevance). Дължината на n-грамата е в интервала [2, 7].

Метод	Кл. алгоритъм	Бр. символи в n-грама в границата на дума	Precision	Recall	micro-F1
Binary Relevance	Multinomial NB	[2, 7]	0.527	**0.513**	0.520
Classifier Chains	SVM	[2, 7]	**0.570**	0.488	**0.526**
Label Powerset	SVM	[2, 6]	0.505	0.498	0.502

Таблица 3 Трансформация на задачата. N-грама от символи в думата.

С n-грама от символи в границата на думата (таблица 3) най-висока прецизност и micro-F1 постига метода верига от класификатори (Classifier Chains), а при оценката за възвращаемост – двоична проверка (Binary Relevance). При тази оптимизация на параметрите, двата най-добре представили се метода използват n-грами с дължина [2, 7].

При оптимизирането на класификаторите методът на опорните вектори (SVM) дава най-добри резултати при метода на всички подмножества от етикети (Label Powerset) и верига от класификатори (Classifier Chains), а за двоична проверка (Binary Relevance) класификаторът с най-високи резултати е Наивен Бейсов класификатор (Multinominal Naive Bayes).

3.7. Адаптирани алгоритми

При използването на адаптираните алгоритми, които са разширения на метода на най-близкия съсед, се използва и оптимизация на параметъра за брой на най-близки съседи в интервала [1, 15].

Метод	k най-близки съседи	Бр. думи в n-грама	Precision	Recall	micro-F1
MLkNN	7	[1, 4]	0.553	**0.385**	0.454
BRkNN	14	[1, 1]	**0.611**	0.382	**0.470**

Таблица 4 Адаптирани алгоритми. N-грама от думи.

Метод	k най-близки съседи	Бр. символи в n-грама	Precision	Recall	micro-F1
MLkNN	8	[3, 6]	0.594	**0.396**	**0.475**
BRkNN	14	[2, 3]	**0.599**	0.354	0.445

Таблица 5 Адаптирани алгоритми. N-грама от символи.

От постигнатите резултати, представени на таблица 4 и таблица 5 алгоритъмът BRkNN постига по-добри резултати за прецизност и micro-F1 при n-думи, както и по-добра възвращаемост и micro-F1 при n-грами, които са

поредица от символи. Но най-добри резултати се постига от MLkNN при n-грама в границата на дума (таблица 6). При избора на параметъра за дължина на n-грамите, когато са поредица от символи в и извън границата на думата, резултатите са подобни.

При оптимизацията на параметъра за брой най-близки съседи - отчетливо се забелязва, че MLkNN работи с по-малък брой съседи, за разлика от BRkNN.

Метод	*k най-близки съседи*	*Бр. символи в n-грама в границата на дума*	*Precision*	*Recall*	*micro-F1*
MLkNN	*5*	*[2, 6]*	***0.612***	***0.395***	***0.480***
BRkNN	*14*	*[2, 3]*	*0.584*	*0.363*	*0.447*

Таблица 6 Адаптирани алгоритми. N-грама от символи в дума.

След анализиране и обобщение на данните от експеримента - методите за трансформация на задачата дават по-добра оценка micro-F1 в сравнение с адаптираните алгоритми, както и по-висока възвращаемост. Но вторите от своя страна постигат сравнително по-висока оценка за прецизност спрямо методите за трансформация на задачата.

Категорично при избор на атрибути и тяхното оптимизиране най-добри резултати се постига при n-грама, която представлява последователност от символи в границите на думата. Доближават се и резултатите като дължина на n-грамите в интервала [2,7].

При трансформацията на задачата - и трите метода постигат сходни резултати като най-висока оценка micro-F1 се постига от метода верига от класификатори (Classifier Chains), като разликата с метода двоична проверка (Binary Relevance) е изключително малка. Отново методът верига от класификатори (Classifier Chains) постига най-висока прецизност, а методът двоична проверка (Binary Relevance) е с най-висока оценка за възвращаемост.

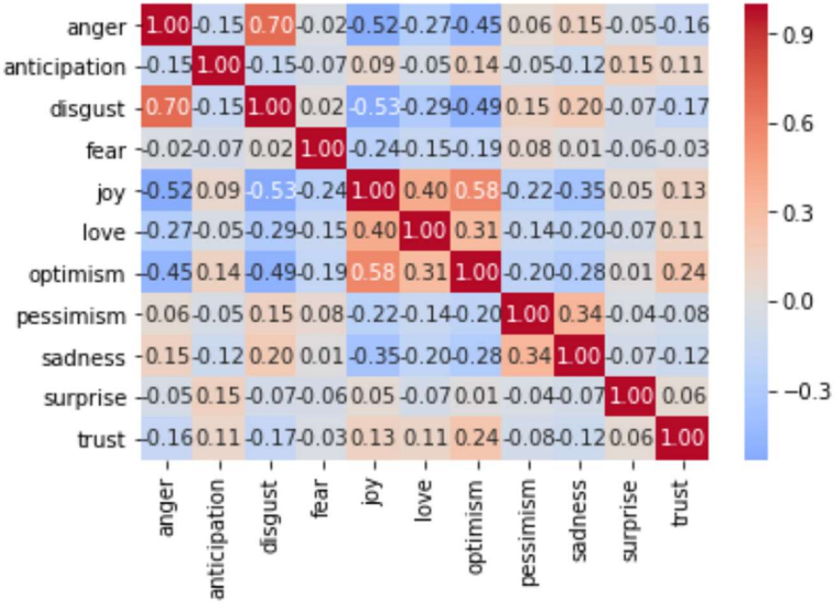
При адаптираните алгоритми MLkNN постига най-високи резултати в сравнение с BRkNN по всички метрики. Оптимизирането на параметри и времето за обучение на MLkNN е сравнително доста по-голямо в сравнение с BRkNN.

3.8. Корелации между емоции

Възможно подобрение при многоетикетна класификация на емоции е използването на знания от предметната област. Такива данни представляват корелационните коефициенти на самите емоции. На фиг.1 са представени извадковите корелационни коефициенти на Пиърсън в данните.

Отчетливо се виждат положително корелиращите емоции “anger” (гняв) и “disgust” (отвращение), “joy” (радост) и “optimism” (оптимизъм), както и емоции, които корелират с отрицателен коефициент - “joy” (радост) и “anger” (гняв), “joy”

(радост) и "disgust" (отвращение). Това означава, че съществува зависимост между всяка двойка емоции. Съществува зависимост и между множества от положителни емоции като „joy" (радост), „love" (любов) и „optimism" (оптимизъм) и отрицателните такива: „anger" (гняв), „fear" (страх) и „disgust" (отвращение).



Фиг. 1 Корелация между емоции в данните за обучение

Тези знания може да се използват както за оптимизиране на избора на класификатори, така и за повишаване точността на предсказване.

Заключение

След направения експеримент и анализирането на резултатите при използването на основните методи за многоетикетната класификация на емоции се наблюдават близки резултати при двете основни групи от методи за решаване на такъв тип задача. Също така класическият и теоретично простият метод двоична проверка (Binary Relevance) дават относително високи оценки и по трите метрики, с които са оценени - прецизност, възвращаемост и micro-F1.

Интерес за бъдещи изследвания представляват корелациите между различните емоции и тяхното използване за по-точната класификация на емоциите в текста.

Библиография

1. Zhi-Hua Zhou, Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. In Pattern Recognition, Volume 40, Issue 7, pages 2038-2048, 2007.
2. S. Aman and S. Szpakowicz. Identifying expressions of emotion in text. In Text, Speech and Dialogue, volume 4629 of LNCS, pages 196–205. Springer, 2007.
3. C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In Proc. HLT–EMNLP, pages 579–586, 2005.
4. C. Yang, K. H.-Y. Lin, and H.-H. Chen. Emotion classification using web blog corpora. In IEEE/WIC/ACM Int'l Conf. on Web Intelligence, pages 275–278, 2007.
5. Mohammad, Saif M., and Felipe Bravo-Marquez. "WASSA-2017 shared task on emotion intensity." *arXiv preprint arXiv:1708.03700,* 2017.
6. Liu, Shuhua Monica, and Jiun-Hung Chen. "A multi-label classification based approach for sentiment classification." *Expert Systems with Applications* 42.3: 1083-1093, 2015.
7. Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. P. Multi-Label Classification of Music into Emotions. In *ISMIR* (Vol. 8, pp. 325-330, 2008.
8. Read, J., Pfahringer, B., Holmes, G., & Frank, E. Classifier chains for multi-label classification. Machine learning, 85(3), 2011.
9. Eleftherios Spyromitros, Grigorios Tsoumakas, Ioannis Vlahavas: An Empirical Study of Lazy Multilabel Classification Algorithms. In: Proc. 5th Hellenic Conference on Artificial Intelligence (SETN 2008), 2008.
10. Kraychev, Boris, and Ivan Koychev. „Automatic Classification of Bulgarian Adjectives on Emotional Semantic Axes. Proceedings of the ERIS, Plovdiv, May, 2014, IMI-BAS & ADIS, pp 121-130, 2014.

AUTOMATIC RECOGNITION OF EMOTIONS IN TEXT

Todor Novakov, Ivan Koychev

***Abstract**: With the rapid development of information technology and the expansion of the Internet, social network users and bloggers generate a huge amount of emotionally rich text. The automatic classification of emotions in the text is an area of interest to both social sciences and applications, such as the discovery of attitudes towards a given product. This paper presents and explores the machine learning methods used to classify emotions in the text. A comparative study of the effectiveness of the classification of the main methods of solving this task has been made. It also presents results from a study of correlations between the main types of emotions in the text.*

***Keywords**: Machine learning, Multi-label classification, Emotion classification*