Abstrakt

Tato práce analyzuje různé metody klasifikace textu za účelem zjištění, zda-li publikované novinové články o konkrétních společnostech umožňují lepší simulaci a predikci volatility akcií dané společnosti. V práci zkoumáme obsah textu publikovaných novinových článků a z toho vycházející sentiment (směr a síla) za použití tří různých přístupů: supervised machine learning Naive Bayes algoritmus, lexicon-based jako zástupce lingvistického přístupu a hybridní Naive Bayes. V rámci hybridního Naive Bayes jsou uvažována pouze slova obsažená v daném lexikonu a nikoliv celý obsah článku. Pro lexicon-based přístup používáme nezávisle dva lexikony, jeden s binárním a jeden vícetřídním hodnocením sentimentu. Sentiment v trénovacím setu pro Naive Bayes byl přiřazen autorem. Z porovnání klasifikační metod založených na machine learning dojdeme k závěru, že všechny metody dosahují podobných výsledků z nichž nejlépe vychází hybridní Naive Bayes používající vícetřídní lexikon. Výstupní kvantitativní data ve formě hodnot sentimentu jsou pak dále zahrnuta do modelování volatility pomocí GARCH. Výsledky ukazují, že informace obsažené v novinových článcích přinášejí další vysvětlující prvek do tradičního GARCH modelu a jsou schopné zlepšit odhad. Nicméně, nejsme schopni získat dost podkladů pro určení nejlepší metody kvantifikace sentimentu. Model používající hybridní Naive Bayes přístup přinesl lepší in-sample výsledky, pro out-of-sample bylo však lepší užít vícetřídní lexikon. Také se nám podařilo ukázat asymetrický efekt, kdy pozitivní i negativní zprávy zvyšují volatilitu, nicméně u zpráv negativních je tento efekt silnější.

Klasifikace JEL	C22, C52, C58, G14, G17, G41
Klíčová slova	volatilita, text, klasifikátor, lexikon, sentiment, novinové články

E-mail autora	`ksenia1105@gmail.com`
E-mail vedoucího práce	`boril.sopov@gmail.com`