Appendix A: Parameters settings

We trained all models using the training data and report the results for the test data. We used full-length texts as input. To transform documents into BoW representations, each dataset was properly pre-processed with word lemmatizing for each word, and all stop words were removed. We also removed all numeric and non-alphabetic characters, as well as frequently occurring words in each corpus, i.e., the top 10 most frequently occurring words in each corpus. When training CWTM, we did not preprocess the documents because it accepts raw text documents as input. When presenting the topics, we preprocessed the topic words generated by CWTM for a fair comparison with other models.

For Gibbs sampling LDA, we set $\alpha = 1/t$ and $\beta = 1/t$, where t is the number of topics for the model and found it outperforms other settings. We trained it over 2000 iterations. For prodLDA and WLDA, we set the batch number to 256 and trained them over 500 epochs for the 20NG, TagMynews and Twitter datasets. We trained them over 10 epochs for the Dbpedia dataset since the number of documents in Dbpedia is much larger. We found no significant difference in the results compared to training them over 7-9 epochs. We used an ADAM optimizer with high momentum $\beta 1 = 0.99$ and a learning rate of 0.002 as it does in Nan et al. (2019). We set the *keep_prob* parameter of ProdLDA to {0.0, 0.2, 0.4, 0.6} and only report the best results from them. We set the Dirichlet prior of WLDA to 0.1 which was suggested by the original paper. The Dirichlet noise of WLDA is set to {0.0, 0.2, 0.4, 0.6} and we report the bests result from them. For CTM, we used the best-performing sentence transformer "all-mpnet-base-v2"[10] to convert each document into a document embedding. We trained it over 100 epochs for the 20NG, TagMynews and Twitter datasets and 2 epochs for the Dbpedia dataset. We tuned the *n_sample* parameter and found that setting it to 40 gives the best results. For CWTM, we built our model on top of "BERT-base-uncased" and we set the epoch number to 20 for the 20NG, TagMynews and Twitter datasets and set it to 1 for the DBpedia dataset. The batch size was set to 16, and the learning rate was set to 1e-3. We also applied a linear scheduler with 10% warmup steps for CWTM. For the soft prompt we used, we set the length of the prompt to 10. We trained each model three times to report the average results.

[10] https://www.sbert.net/docs/pretrained_models.html

Appendix B: Document classification accuracy across different settings

We plot the classification performance of different models with different numbers of topics settings in Figure B1. The number of topics is set to Z= {10, 20, 50, 100, 200}. CWTM achieved better performance than the other models on the TagMyNews, Twitter, Dbpedia, and AGNews datasets. Although LDA outperforms CWTM on the 20NG dataset, CWTM has better performance than the other neural topic model baselines.

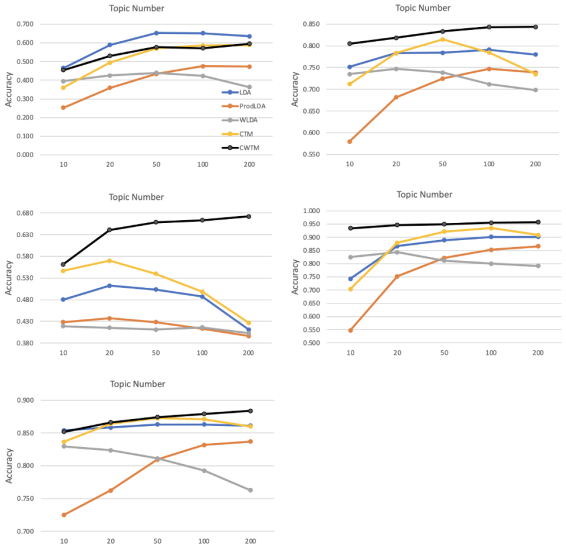

Figure B1: Document classification accuracy across different numbers of topics by different models for 20NG (top row left); TagMyNews (top row right), Twitter (2nd row left), DBpedia (2nd row right), and AGNews (bottom row).

Appendix C: Topic words from CWTM on different datasets

Topic	Top 10 words
topic 1	key, chip, radio, circuit, phone, power, battery, encryption, signal, line
topic 2	car, bike, engine, mile, brake, tire, wheel, rear, dealer, bmw
topic 3	god, christian, bible, jesus, scripture, christ, church, sin, lord, holy
topic 4	space, nasa, orbit, doctor, patient, disease, medical, launch, drug, shuttle
topic 5	game, player, hockey, team, baseball, playoff, pitcher, nhl, play, goal
topic 6	drive, card, driver, disk, bus, memory, ram, controller, mac, machine
topic 7	argument, agree, moral, opinion, discussion, question, point, assume, belief, fact
topic 8	armenian, gun, child, fbi, weapon, police, arab, government, law, jew
topic 9	window, program, application, post, unix, code, software, command, server, run
topic 10	test, david, ditto, stuff, cheer, hey, deleted, tony, sex, michael

Table 8: Topic words from CWTM on the 20NG dataset.

Topic	Top 10 words
topic 1	nuclear, earthquake, tsunami, plant, tornado, quake, radiation, power, oil, reactor
topic 2	stock, bank, billion, investor, market, fund, price, financial, share, debt
topic 3	yankee, inning, sox, phillies, mets, run, hitter, pitcher, jay, baseball
topic 4	broadway, theater, musical, idol, tony, revival, play, show, stage, star
topic 5	study, cancer, risk, drug, diabetes, patient, disease, woman, heart, treatment
topic 6	open, round, federer, murray, nadal, french, andy, djokovic, final, master
topic 7	trial, court, case, police, accused, charge, prosecutor, judge, charged, guilty
topic 8	apple, google, microsoft, android, tablet, phone, ipad, browser, service, window
topic 9	connecticut, east, kentucky, ncaa, butler, ohio, duke, notre, big, texas
topic 10	rebel, libyan, force, libya, gaddafi, nato, government, japan, military, syrian

Table 9: Topic words from CWTM on the TagMyNews dataset.

Topic	Top 10 words
topic 1	customer, twitter, ur, tweet, email, account, post, awe, service, app
topic 2	fucking, fuck, shit, guy, hilarious, hell, pout, man, funny, team
topic 3	terrorism, anger, outrage, terror, black, rage, pakistan, angry, fuming, trump
topic 4	amazing, blue, lively, news, music, rock, musically, tonight, great, free
topic 5	hilarious, horrible, fuming, fury, depressing, rage, furious, bitter, irritate, nightmare
topic 6	good, happy, lost, year, work, back, night, home, tomorrow, wait
topic 7	depression, fear, sadness, anxiety, afraid, terrible, bad, worst, hate, cry
topic 8	today, phone, guy, year, work, service, back, thing, im, call
topic 9	live, broadcast, ly, gbbo, lol, birthday, gt, snap, wow, snapchat
topic 10	watch, musically, blue, horror, nightmare, absolutely, delight, exhilarating, glee, playing

Table 10: Topic words from CWTM on the Twitter dataset.

Topic	Top 10 words
topic 1	aircraft, car, engine, wing, seat, locomotive, fighter, model, motorcycle, air
topic 2	journal, university, newspaper, published, editor, peer, reviewed, academic, college, magazine
topic 3	high, public, grade, student, pennsylvania, township, secondary, education, historic, house
topic 4	football, played, footballer, player, play, professional, league, hockey, playing, team
topic 5	directed, starring, drama, book, movie, comedy, star, story, written, thriller
topic 6	navy, ship, class, launched, hm, commissioned, submarine, laid, vessel, built
topic 7	band, record, studio, rock, song, released, singer, track, single, ep
topic 8	mountain, peak, river, range, lake, mount, alp, elevation, summit, flow
topic 9	plant, found, genus, moth, endemic, habitat, flowering, native, tree, grows
topic 10	persian, iran, romanized, rural, gmina, poland, population, province, village, voivodeship

Table 11: Topic words from CWTM on the Dbpedia dataset.

Topic	Top 10 words
topic 1	stock, price, dollar, market, investor, higher, future, high, rise, rose
topic 2	court, charge, case, police, arrested, charged, lawsuit, judge, trial, drug
topic 3	oil, space, moon, spacecraft, crude, nasa, titan, station, saturn, planet
topic 4	arsenal, goal, liverpool, chelsea, united, manchester, striker, england, test, champion
topic 5	corp, million, billion, deal, buy, bid, group, business, takeover, share
topic 6	inning, sox, yankee, league, giant, astros, dodger, cub, twin, nl
topic 7	music, game, store, video, dvd, nintendo, sony, make, song, ipod
topic 8	baghdad, militant, iraqi, iraq, insurgent, rebel, killed, soldier, attack, troop
topic 9	microsoft, linux, software, version, window, source, server, system, application, enterprise
topic 10	open, round, final, championship, master, cup, champion, federer, seed, hewitt

Table 12: Topic words from CWTM on the AGNews dataset.